Structural Similarities Between Language Models and Neural Response Measurements

Jiaang Li[*] JLI@HUM.KU.DK
Copenhagen University

Antonia Karamolegkou[*] ANTKA@DI.KU.DK
Copenhagen University

Yova Kementchedjhieva YOVA@DI.KU.DK
Copenhagen University

Mostafa Abdou MA4231@PRINCETON.EDU
Princeton University

Sune Lehmann SLJO@DTU.DK
Technical University of Denmark, Pioneer Centre for AI, Denmark

Anders Søgaard SOEGAARD@DI.KU.DK
Copenhagen University, Pioneer Centre for AI, Center for Philosophy of AI, Denmark

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane.

Abstract

Large language models have complicated internal dynamics, but induce representations of words and phrases whose geometry we can study. Human language processing is also opaque, but neural response measurements can provide (noisy) recordings of activations during listening or reading, from which we can extract similar representations of words and phrases. Here we study the extent to which the geometries induced by these representations, share similarities in the context of brain decoding. We find that the larger neural language models get, the more their representations are structurally similar to neural response measurements from brain imaging.

Keywords: Neural Representations, Language Representations, Brain Decoding

1. Introduction

Understanding how the brain works has intrigued researchers for many years. This challenge has given rise to the field of brain decoding, where the goal is to interpret the information encoded in the brain while a person is engaged in a specific cognitive task, such as reading or listening to language. By analyzing representations of neural activity across different brain regions, researchers can develop computational models that link specific patterns of brain activity to linguistic elements, such as words or sentences. This direction of research opens avenues for advancing our understanding of neurological disorders, developing innovative treatments, and enhancing the quality of life for individuals with disorders.

[*] Equal Contribution

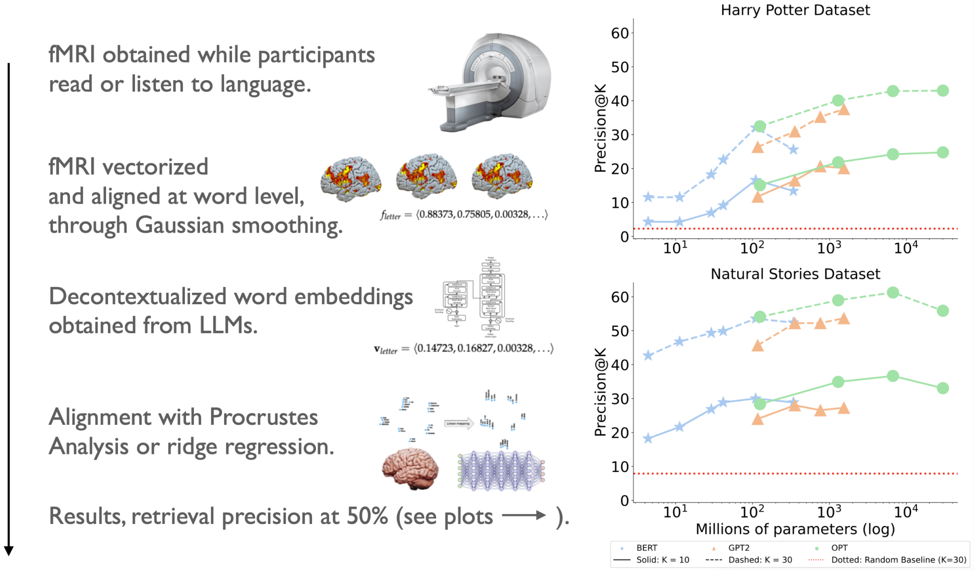

Figure 1: **Experimental flow and main results.** We run experiments with three families of LLMs (comparing LLMs of different sizes within families), two fMRI datasets, and three projection algorithms, and results are the same across all combinations: LLMs converge toward human-like representations, enabling (P@10) retrieval rates of up to 25%, i.e., a quarter of all concepts can be decoded from the fMRI signals. The datasets, our Gaussian smoothing technique, and the projection methods are described in §3. Right side: **Convergence results for three families of LLMs across two datasets, using Procrustes analysis**. Convergence is consistent, and some retrieval rates are remarkably high, decoding almost half of the words correctly to a neighborhood of 30 word forms, which surpasses the random retrieval baselines represented by the dotted red lines.

In this paper, we investigate the alignment between the representations of words in LLMs and the neural response patterns observed in the human brain during language processing. What emerges is a striking structural similarity between these two sets of representations, manifesting as a geometric congruence in high-dimensional vector spaces. To quantify this alignment, we employ rigorous evaluation methods, including ridge regression, representational similarity analysis (RSA) (Kriegeskorte et al., 2008), and Procrustes analysis (Kementchedjhieva et al., 2018) (if $d = d'$). These methodologies enable us to quantify the extent of isomorphism between LLMs and neural responses, e.g., functional magnetic resonance imaging (fMRI). Figure 1 illustrates the experimental flow and main results.

It is a common practice to evaluate LLMs by measuring their performance on benchmark data and protocols (Lewkowycz et al., 2022; Mitchell and Krakauer, 2023). Doing so is aimed to infer what LLMs have learned, from how they behave. The methodology is behaviorist and has obvious limitations. We instead suggest exploring the inside of LLMs and our brains – or, to be precise, their representational *geometries*. Our investigations span various

LLM families, word embeddings, and diverse datasets, consistently revealing high degrees of structural congruence. Our main contributions are as follows:

- We find a remarkable structural similarity between how words are represented in LLMs, and the neural response measurements of humans reading the same words. The LLM representations of a vocabulary form a geometry in a d-dimensional vector space; and the neural response measurements from one or more participants reading these words in a brain scanner, form in a similar way a geometry in a d'-dimensional space.
- We present experiments for three families of LLMs (as well as one static embedding method), with two different fMRI datasets, and three evaluation methods (ridge regression, RSA, and Procrustes analysis) to compute the structural similarity (degree of isomorphism) between these two modal geometries.
- Across the board, we see high degrees of isomorphism, enabling decoding or retrieval performance (precision-at-k, a.k.a P@k) of up to P@10≈25% (with random performance being P@10<1%). Word-level brain decoding thus seems feasible particularly as language models increase in size.

2. Related Work

Over the past decade, researchers have explored the relationship between neural and language representations by predicting text from brain activity (Søgaard, 2016; Ramakrishnan and Deniz, 2021; Fereidooni et al., 2020).Pereira et al. (2018) were the first to build regression models to predict sentence representations from brain scans. Extending this work, Minnema and Herbelot (2019) investigate several metrics to evaluate the decoder performance. Apart from regression, Sun et al. (2019) also use similarity-based decoders where the decoder is trained to map brain images to distinct sentence representations in both structured and unstructured settings. Affolter et al. (2020) use a neural network model to facilitate the brain-to-word regression decoder, and evaluate on unseen subjects for a more realistic approach. Oota et al. (2022) propose two novel setups using multi-view and cross-view regression decoders that predict semantic concepts and vector representations respectively. Zou et al. (2022) suggest a neural decoder in a cross-modal cloze setting predicting the target word given a contextual prompt. Finally, Tang et al. (2022) build a decoder that reconstructs continuous language instead of individual words or sentences. Our focus is not on building a better decoding system, but rather on exploring the alignment between neural and language representational spaces within such systems.

3. Methodology

We begin with a description of the tools and methods we used to align language model representations with fMRI recordings. We experiment with three families of LLMs, comparing the word representations they induce to human representations obtained from two different neural response measurement datasets. We use three different comparison methods, leading to a total of 18 experiments, which all confirm the same trend. For further information regarding our models and their configuration, refer to the Appendix.

3.1. Data description and pre-processing

3.1.1. fMRI datasets

fMRI is a non-invasive neural response measurement technique that records on a spatial resolution in the region of 1 to 6 millimeters, higher than any other technique. fMRI records activity (blood flow) in the entire network of brain areas engaged when subjects undertake particular tasks. On the downside, fMRI is somewhat susceptible to influences of non-neural changes, and the temporal response is poor relative to the electrical signals that define neuronal communication. To compensate for low temporal resolution, we introduce Gaussian smoothing below. The datasets are: Harry Potter Dataset (Wehbe et al., 2014) (8 subjects), and Natural Stories Audio Dataset (Zhang et al., 2020) (19 subjects). Both datasets are publicly available.

3.1.2. Gaussian smoothing

We use Gaussian smoothing to extract word-level neural response measurements in our two datasets. Gaussian smoothing has been used before to study speech-aligned fMRI data (Bingel et al., 2016; Brodoehl et al., 2020). In cases where fMRI data is not collected at the granularity of individual words, we can use Gaussian smoothing to generate word-level fMRI information. For instance, to obtain the fMRI vector for a specific word like ”Harry” at a given time point t (Harry$_t$), we can extract the fMRI vectors for a certain timeframe T around t, such as $t \pm T$ seconds. We then apply Gaussian smoothing to this set of vectors, resulting in a final vector that represents the fMRI information for the word ”Harry$_t$”. This approach has potential benefits for fMRI analysis in various applications, such as studies of language processing and cognitive neuroscience. By generating word-level fMRI information using Gaussian smoothing, we can potentially extend the scope from sequence-level to word-level, and improve the interpretability and accuracy of the results. Extracting word-level signals differentiates our work from most of the previous, and is shown to be crucial in recent work on brain decoding (Tang et al., 2022).

3.2. Models

3.2.1. Non-auto-regressive models

Non-auto-regressive models are a type of machine learning models that take in an input sequence of text and generate a single output vector representation for the entire sequence. During training, some words are masked and the model learns to predict the masked words based on context. We use the BERT (Devlin et al., 2019) family of language models as an example of non-auto-regressive models.

3.2.2. Auto-regressive models

Auto-regressive models generate output sequences by predicting each element in the sequence based on the previously generated elements. In other words, the output is generated one element at a time, with the model conditioned on the previous output elements. These language models are used to generate text but typically provide slightly worse similarity estimates. We use two auto-regressive language model families: GPT2 (Radford et al., 2019) and OPT (Zhang et al., 2022).

3.3. Comparison and projection methods

3.3.1. Representational Similarity Analysis

Relational Similarity Analysis (RSA) is a multivariate analysis technique commonly used in cognitive neuroscience and computational linguistics to compare the similarity between two sets of representations (Kriegeskorte et al., 2008). RSA can be used to measure the similarity between the neural activity patterns observed in the fMRI data and the representations learned by LLMs. RSA operates by first representing the neural activity and language model features as vectors in a high-dimensional space. The similarity between these vectors is then quantified using a rank-based correlation metric. We perform RSA following (Lepori and McCoy, 2020). Let X and Y be two sets of representations. We calculate their representational dissimilarity matrices (RDMs) as $\mathbf{D}_X$ and $\mathbf{D}_Y$, respectively[1]. We then compare the representational geometries using Spearman's rank correlation coefficient, denoted as $\rho(\mathbf{D}_X, \mathbf{D}_Y)$.

3.3.2. Ridge regression

Ridge regression is a widely used method in statistics and machine learning to address the issue of multicollinearity, which can arise when there are highly correlated predictor variables in a linear regression model. In contrast to Toneva and Wehbe (2019), who utilized ridge regression for encoding fMRI data, our approach focuses on decoding, i.e. predicting language from fMRI. We achieve this by establishing a model that captures the connection between brain signals and individual dimensions within the language model representations. The models are trained to predict the signal of word w in layer l, denoted as y_{l^w}, using the vector of fMRI voxels for that word, x^w. For each subject and layer l, we employ cross-validation to estimate the predictiveness of the fMRI representation of the word in each dimension i. In each fold, the fMRI data matrix with total n dimension denoted as $X = x_{w^1}, x_{w^2}, ..., x_{w^n}$, and the semantic vector matrix with m dimension, denoted as $Z = z_{w^1}, z_{w^2}, ..., z_{w^m}$, are split into corresponding training and validation matrices which are individually normalized to have a mean of 0 and a standard deviation of 1 for each dimension across words, ending with training matrices X^R and $Z^{R,l}$, as well as validation matrices X^V and $Z^{V,l}$. Using the training fold, we estimate a model $\theta^{i,l}$ as follows:

$$\arg\min_{\theta^{i,l}} ||z^{R,i} - X^R \theta^{i,l}||_2^2 + \lambda^i ||\theta^{i,l}||_2^2$$

To identify the best λ^i for each dimension i that minimizes the nested cross-validation error, we employ a ten-fold nested cross-validation. Subsequently, we estimate $\theta^{i,l}$ using λ^i on the entire training fold. Thus, the predictions for each dimension in the validation fold are obtained as $p^l = X^V \theta^{i,l}$.

3.3.3. Procrustes Analysis

We use Procrustes Analysis, a form of statistical shape analysis, to align brain fMRI representations with those of language models, using a bimodal dictionary. Procrustes Analysis

[1] The code we used was taken from: `https://github.com/mlepori1/Picking_BERTs_Brain`

Datasets	U.W.	P@1	P@5	P@10	P@30	P@50	P@100
Harry Potter$_{Random}$	1291	0.08%	0.39%	0.77%	2.32%	3.87%	7.74%
Harry Potter$_{FastText}$		0.36%	3.66%	6.43%	12.76%	17.22%	26.52%
Natural Stories$_{Random}$	381	0.26%	1.31%	2.62%	7.87%	13.12%	26.25%
Natural Stories$_{FastText}$		0.00%	1.88%	5.62%	17.27%	24.24%	39.89%

Table 1: Two different P@k baselines with $k \in \{1, 5, 10, 30, 50, 100\}$ of two datasets. The random retrieval baselines are calculated by the U.W. in stimulus content, respectively. U.W. = the number of unique words.

is a method for matching corresponding points in two shapes and finding the transformation (translation, rotation, and scaling) that best aligns them. Specifically, we seek to find the orthogonal matrix Ω that optimally maps the brain fMRI matrix A representing brain responses to the words onto the language model matrix B, i.e. the language model representations of the words, using the $\min_R |R - M|_F$ problem subject to $R^T R = I$, which can be solved using singular value decomposition with $R = UV^T$.

4. Experimental Setup

4.1. fMRI-text Dictionary Complementation

We build a bimodal dictionary that associates fMRI data with corresponding textual information. Considering the context in which words are presented, it becomes evident that the brain's response to a particular word may vary significantly across different sentences. This dynamic response suggests that, within our constructed dictionary, the relationship between fMRI recordings and textual entries exhibits a many-to-one correspondence. We employ a four-fold cross-validation approach that takes into account unique words, thereby preventing any potential train-test leakage. Due to individual differences among subjects, our experiments are conducted based on each subject's responses. We report the averaged results across all subjects.

4.2. Evaluation - Linear Projection

To assess the effectiveness of regression and alignment techniques, we employ the P@k metric, which quantifies the ratio of accurate predictions within the top k predictions. This evaluation metric offers a more cautious and robust assessment (Karamolegkou et al., 2023). For Procrustes analysis, we induce it from a small set of point pairs and test it on held-out data measuring the P@k (Lample et al., 2018), whereas for Regression we use all point pairs. Ensuring consistent dimensionality between the source and target spaces is a crucial prerequisite for successful alignment. In instances where a dimensionality mismatch arises, we employ principal component analysis to reduce the dimensionality of the larger space. To find the top k predictions we use Cross-domain similarity local scaling (CSLS). This method is often used to evaluate the similarity between mapped source and target words and is an improvement of the traditional Nearest Neighbor (NN) methods Lample et al. (2018). See more details for this metric in the Appendix.

Random retrieval baseline. P@k is a metric that quantifies the proportion of words for which the LLM's representation serves as one of the k-nearest neighbors to the corresponding fMRI encoding. In essence, word-level decoding involves a straightforward nearest-neighbor retrieval process within the projected space. It's crucial to note that our target vector space, which represents the language model, contains hundreds of vectors. This feature sets our random baseline P@1 $< 0.1\%$. Our target space of the text material in fMRI datasets makes the random retrieval baseline: P@1 $= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{U} \times 100\%$, where N represents the total number of unique words; i iterates over all words in the material; U refers to the total number of unique words.

FastText baseline. Surpassing the random baseline does not necessarily establish that LLMs are inherently more aligned with fMRI measurements. To address this concern, we conduct a secondary baseline alignment experiment by aligning fMRI recordings with word representations from fastText Bojanowski et al. (2017). Further details can be found in Table 1. In practical applications, our mappings exhibit significantly higher precision, reflecting the structural similarities between the language model and human brains.

5. Results & Discussion

5.1. Main Results.

Our main results are presented in Figure 2 which illustrates the averaged results across all subjects, and the convergence of three families of LLMs on representations that are remarkably similar to those seen in neural response measurements. These results are consistent across two fMRI datasets and three mapping methods. See Appendix D (Figure 5) for similar results with RSA. The scores are plotted by model size, showing the convergence toward brain-like representations as LLMs increase in size. The best scores indicate that LLMs up to 1.5B parameters can achieve alignments such that a bit more than 1 in 5 words are decoded correctly,[2] and a bit more than 2 in 5 almost correctly (within neighborhoods of 20-30 word forms). To gain a qualitative sense of the alignment between brain signal and LLMs representations, see Figure 3. The results are obtained with limited supervision for learning the mapping. In fact, we only rely on 950 data points to induce this linear projection, a small number given the high dimensionality of the derived word representations; see §3 for details.

5.2. Discussion

Our findings reveal a strong similarity between language model word representations and human brain responses to language stimuli. As these neural language models expand in size, their representations become more akin to the patterns observed in neural responses from the fMRI scans. This discovery points to the development of human-like representations within these large-scale language models, offering valuable insights into the intricate relationship between artificial intelligence and human cognitive processes.

[2] The reason we count P@5 or P@10 as correct decoding is that a neighborhood of 5-10 words will tend to consist of inflections of the same lemma or synonymous words (Kementchedjhieva et al., 2019). P@1 would amount to guessing the lemma, the exact inflection, and the correct spelling variant.

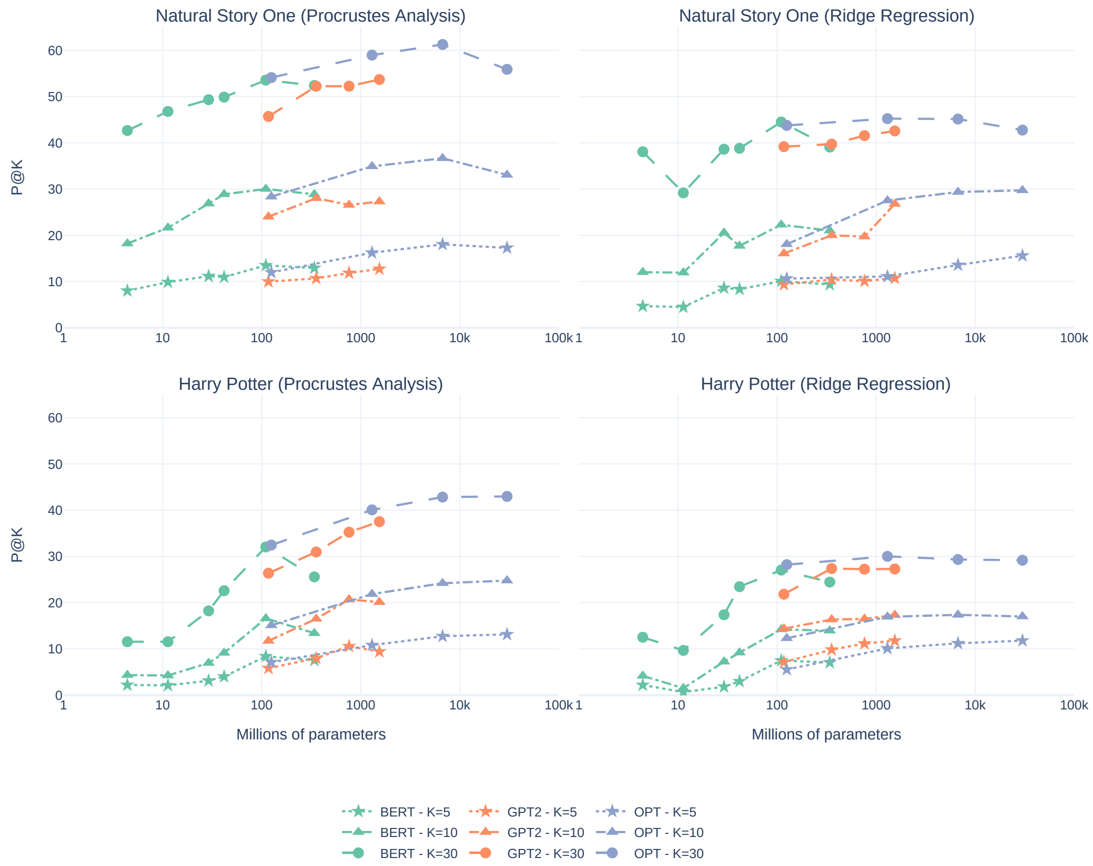

Figure 2: **Convergence Results for Three Families of Language Models on Two Datasets**. The task for the Harry Potter dataset here is: Given a neural response, which word in a vocabulary of 1,291 words, was read at the time the response was recorded? Random retrieval baseline P@10 is *less* than 1%, while the FastText baseline P@10 is *less* than 7%. See more details of baselines in Table 1.

Newman's objection? Philosophers argue whether structural similarities (isomorphisms, homomorphisms, etc.) between representations and what is represented, are sufficient for content (Shea, 2007; Mollo and Millière, 2023). Their concerns have their origin in Newman's objection to Russellian structural realism (Newman, 1928). Briefly put, Newman showed that structuralist descriptions that abstract away from all but the logical structure, and simply assert the existence of a relation that induces a graph isomorphism between the representation, and what is represented, are indeed trivial. Any LLM will, in other words, induce word representations such that the nearest neighbor graph over the word vocabulary $\mathcal{V}$ such that there exists a relation that is isomorphic to that graph. Mollo and Millière (2023), for example, bring up Newman's objection and write:

> philosophical work on theories of representational content has long established [. . . that m]orphisms between two sets of objects or properties are trivial to find, and rely solely on the existence of morphisms between internal representations

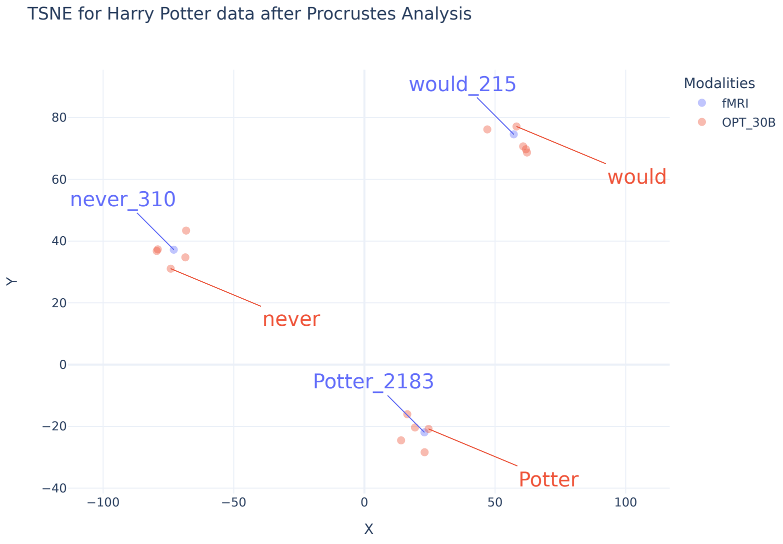

Figure 3: **t-SNE plot of fMRI and LLM representations** using OPT-30B (large, uncased) over selected target words from the Harry Potter dataset. Here we visualize the P@5, i.e. the top 5 predictions for the fMRI representation. The fMRI representation '$\texttt{Potter}_{2183}$' has the LLM representation *Potter* among its 5-nearest neighbors. The fMRI representation associated with '$\texttt{never}_{313}$' is not as close as the LLM representation *never* – but still with *never* as the top-5 guess. That said, the words *Potter* and *would* are decoded correctly by our alignment (top-1 guess or P@1).

> and structured domains in the world could lead to a trivialization of the notions of representation and meaning . . .

However, Newman's objection only holds if all there is posited is the existence of *some* relation. If the relations are properly restricted, isomorphism is far from trivial. One observation that goes all the way back to Carnap's *Aufbau* (Carnap, 1967)[3] is: Structural similarities are generally trivial to obtain, but if the relations (distances in the vector space) serve a purpose (do work for the system), structural similarities can ground content. Structural similarity is evidently sufficient to solve semantic problems, such as bilingual dictionary induction (Søgaard et al., 2019) or multi-modal alignment (Li et al., 2023). The fact that fMRI vectors exhibit structural similarities to LLMs (and by transitivity, across languages and to computer vision models), is suggestive of such similarities playing a role in grounding.

In our case, we are not simply positing an isomorphic relation in neural responses. We are positing an isomorphism between two very specific relations: the nearest neighbor graph in the LLM representations, and the nearest neighbor graph in the fMRI data. In fact, these two relations are the *same* relation, something which Newman himself proposed as a remedy to his own objection. It should thus be clear that the result presented here is far from trivial.

[3] Russell arguably had a similar response (Pashby, 2015).

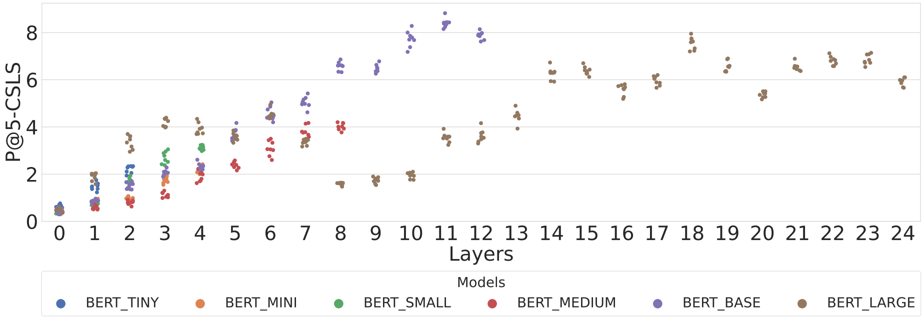

Figure 4: **Alignment precision results across layers with Harry Potter dataset.** The plot shows alignment with fMRI improves with model depth, for BERT and Procrustes Analysis; see Appendix D for similar plots for other LLMs and projection methods.

Where are LLMs most brain-like? We also consider at what layers the different language models align best with the representations extracted from the fMRI data. The results presented in Figure 4 and the Appendix D are unambiguous and show that deeper representations align better with neural response measurements. This holds across all architectures and model sizes. Interestingly, the alignment improvements at deeper layers do not wear off to reach a plateau. Our results, in fact, suggest that better alignment results can be achieved by training even deeper models. This may also explain the strong correlation between depth and generalization often observed in the literature (Goodfellow et al., 2014). It has generally been found that the inner-most layers in LLMs encode for syntax, whereas the outer layers encode for semantics and pragmatics. One way to understand our results is therefore that similarities between representations in human brains and LLMs are predominantly driven by semantics and pragmatics.

6. Conclusion

We presented a series of experiments showing that across three families of language models, word representations converge toward being structurally similar to human neural responses. The larger and better the language models get, the more their representations align with human representations. This result holds across datasets and three evaluation methods. We have discussed the philosophical significance of this result, including why Newman's objection does not apply. We include a discussion of the limitations of our work and provide an ethical statement in the Appendix.

Acknowledgments Thanks to the anonymous reviewers for their helpful feedback. This work is supported by the Novo Nordisk Foundation (grant NNF 20SA0066568). Jiaang Li is supported by Carlsberg Research Foundation (grant CF22-1432). Also, Antonia Karamolegkou is supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023'.

References

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: Decoding brain activity for language generation, 2020.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *CoRR*, abs/2001.08435, 2020. URL https://arxiv.org/abs/2001.08435.

Joachim Bingel, Maria Barrett, and Anders Søgaard. Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1071. URL https://aclanthology.org/P16-1071.

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. ISSN 2307-387X.

Stefan Brodoehl, Christian Gaser, Robert Dahnke, Otto W. Witte, and Carsten M. Klingner. Surface-based analysis increases the specificity of cortical activation patterns and connectivity results. *Scientific Reports*, 10, 2020.

Rudolf Carnap. *The Logical Structure of the World*. Berkeley: University of California Press, 1967.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*, 2018. URL http://arxiv.org/abs/1710.04087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Sam Fereidooni, Viola Mocz, Dragomir Radev, and Marvin Chun. Understanding and improving word embeddings through a neuroscientific lens. *bioRxiv*, pages 2020–09, 2020.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. 2014. URL https://arxiv.org/pdf/1312.6082.pdf.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping brains with language models: A survey, 2023.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1021. URL https://aclanthology.org/K18-1021.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1328. URL https://aclanthology.org/D19-1328.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, 2:4, November 2008.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Michael Lepori and R. Thomas McCoy. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3637–3651, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.325. URL https://aclanthology.org/2020.coling-main.325.

Aitor Lewkowycz, Ambrose Slone, Anders Andreassen, Daniel Freeman, Ethan S Dyer, Gaurav Mishra, Guy Gur-Ari, Jaehoon Lee, Jascha Sohl-dickstein, Kristen Chiafullo, Liam B. Fedus, Noah Fiedel, Rosanne Liu, Vedant Misra, and Vinay Venkatesh Ramasesh. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Technical report, 2022.

Jiaang Li, Yova Kementchedjhieva, and Anders Søgaard. Implications of the convergence of language and vision model geometries. *arXiv preprint arXiv:2302.06555*, 2023.

Gosse Minnema and Aurélie Herbelot. From brain space to distributional space: The perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2021. URL https://aclanthology.org/P19-2021.

Melanie Mitchell and David C. Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120,

2023. doi: 10.1073/pnas.2215907120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2215907120.

Dimitri Coelho Mollo and Raphaël Millière. The vector grounding problem, 2023.

M. H. A. Newman. Mr. russell's causal theory of perception. *Mind*, 37(146):26–43, 1928. doi: 10.1093/mind/xxxvii.146.137.

Subba Reddy Oota, Jashn Arora, Manish Gupta, and Raju S. Bapi. Multi-view and cross-view brain decoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 105–115, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.10.

Thomas Pashby. Understanding russell's response to newman. 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9, 03 2018. doi: 10.1038/s41467-018-03068-4.

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

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531, dec 2010. ISSN 1532-4435.

Kalyan Ramakrishnan and Fatma Deniz. Non-complementarity of information in word-embedding and brain representations in distinguishing between concrete and abstract words. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–11, 2021.

Nicholas Shea. Content and its vehicles in connectionist systems. *Mind and Language*, 22 (3):246–269, 2007. doi: 10.1111/j.1468-0017.2007.00308.x.

Anders Søgaard. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2521. URL https://aclanthology.org/W16-2521.

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. *Cross-Lingual Word Embeddings*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States, 2 edition, 2019. doi: 10.2200/S00920ED2V01Y201904HLT042.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33017047. URL https://doi.org/10.1609/aaai.v33i01.33017047.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *bioRxiv*, 2022. doi: 10.1101/2022.09.29.509744. URL https://www.biorxiv.org/content/early/2022/09/29/2022.09.29.509744.

Mariya Toneva and Leila Wehbe. *Interpreting and Improving Natural-Language Processing (in Machines) with Natural Language-Processing (in the Brain)*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847, 2018.

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, November 2014.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *bioRxiv*, 2020. doi: 10.1101/649939. URL https://www.biorxiv.org/content/early/2020/03/26/649939.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Cross-modal cloze task: A new task to brain-to-word decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 648–657, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.54. URL https://aclanthology.org/2022.findings-acl.54.

Appendix A. Limitations

Our study demonstrates the precise mapping of neural response measurements to language model representation spaces through supervised learning. However, our findings are subject to certain constraints. The utilization of fMRI signals with limited temporal resolution, albeit partially mitigated through Gaussian smoothing, may introduce potential confounds. Additionally, our primary focus on the English language narrows the generalization ability of our results to languages with different linguistic structures. Furthermore, relying on a single participant for each alignment may introduce individual variability that could influence our conclusions. Moreover, our paper emphasizes the philosophical interpretation of the linear mapping results, leaving the technical aspects of this alignment largely unexplored. To ensure the robustness and broader applicability of our findings, future research should encompass diverse languages, and participant groups, and delve deeper into the technical underpinnings of the observed alignment between neural responses and language model representations.

Appendix B. Ethics

In our research, we analyze two publicly available fMRI datasets (Harry Potter Dataset and Natural Stories Dataset). We did not collect any new dataset for our study. We encourage readers to refer to the terms of use provided by the respective dataset sources for a more comprehensive understanding of their ethical guidelines and data usage policies. We do not foresee any harmful uses of this line of research that compares representational spaces.

Appendix C. Implementation

Our implementation is based on PyTorch v.1.13.1 (Paszke et al., 2019) and Transformer v4.25.1 (Wolf et al., 2020) for Python 3.9.13 and builds on code from the repositories in Table 2. You can find details for our models in Table 3. We took the pretrained models without fine-tuning them.

Cross-domain Similarity Local Scaling (CSLS). A method used often to evaluate different word representations is Nearest Neighbors (NN). Nearest neighbors are naturally asymmetric, which means if y is a K-NN of x, it does not follow that x is also a K-NN of y. In high-dimensional spaces (Radovanović et al., 2010), the nearest neighbor rule can

LMs	Links
BERT_TINY BERT_MINI BERT_SMALL BERT_MEDIUM BERT_BASE BERT_LARGE	https://huggingface.co/google/bert_uncased_L-2_H-128_A-2 https://huggingface.co/google/bert_uncased_L-4_H-256_A-4 https://huggingface.co/google/bert_uncased_L-4_H-512_A-8 https://huggingface.co/google/bert_uncased_L-8_H-512_A-8 https://huggingface.co/bert-base-uncased https://huggingface.co/bert-large-uncased
GPT2_BASE GPT2_MEDIUM GPT2_LARGE GPT2_XL	https://huggingface.co/gpt2 https://huggingface.co/gpt2-medium https://huggingface.co/gpt2-large https://huggingface.co/gpt2-xl
OPT_125M OPT_1.3B OPT_6.7B OPT_30B	https://huggingface.co/facebook/opt-125m https://huggingface.co/facebook/opt-1.3b https://huggingface.co/facebook/opt-6.7b https://huggingface.co/facebook/opt-30b

Table 2: Links of 14 Transformer-based language models used in our experiments.

LMs	Hidden Layers	Hidden Size	Attention Heads	Total # of Params	Datasets
BERT	2	128	2	4.4M	BooksCorpus (Zhu et al., 2015), English Wikipedia (Devlin et al., 2019)
	4	256	4	11.3M	
	4	512	8	29.1M	
	8	512	8	41.7M	
	12	768	12	110.1M	
	24	1,024	16	336M	
GPT2	12	768	12	117M	WebText (Radford et al., 2019)
	24	1,024	16	345M	
	36	1,280	20	762M	
	48	1,600	25	1,542M	
OPT	12	768	12	125M	BooksCorpus, CC-Stories(Trinh and Le, 2018), CCNewsV2(Zhang et al., 2022), The Pile(Gao et al., 2021), Pushshift.io Reddit dataset (Baumgartner et al., 2020)
	24	1,024	32	1.3B	
	32	4,096	32	6.7B	
	48	7,168	56	30B	

Table 3: The 14 language models used in our experiments. Table 2 lists the links of LMs.

lead to a phenomenon called hubness, where some vectors (hubs) are nearest neighbors of many other points, while others (anti-hubs) are not nearest neighbors of any point. This is detrimental to matching pairs based on the nearest neighbor rule. To address this issue, Conneau et al. propose a bi-partite neighborhood graph (Conneau et al., 2018), in which each word of a given dictionary is connected to its K nearest neighbors in the other language. The neighborhood of a mapped source word embedding Wx_s, denoted as $N_T(Wx_s)$, is represented on the bipartite graph. It consists of K elements, all of which are words from the target language. Similarly, the neighborhood of a target word t, denoted as $N_S(y_t)$. The mean similarity between a source embedding x_s and its corresponding target neighborhood is considered as:

$$r_T(Wx_s) = \frac{1}{K} \sum_{y_t \in N_T(Wx_s)} cos(Wx_s, y_t),$$

where cos(.,.) means cosine similarity. They use $r_S(y_t)$ to represent the mean similarity of a target word y_t to its neighborhood. The definition of Cross-domain Similarity Local Scaling (CSLS) between mapped source words and target words is

$$CSLS(Wx_s, y_t) = 2cos(Wx_s, y_t) - r_T(Wx_s) - r_S(y_t).$$

Appendix D. More Results

We provide more experimental results to enhance our main discussions/findings described in the main paper.

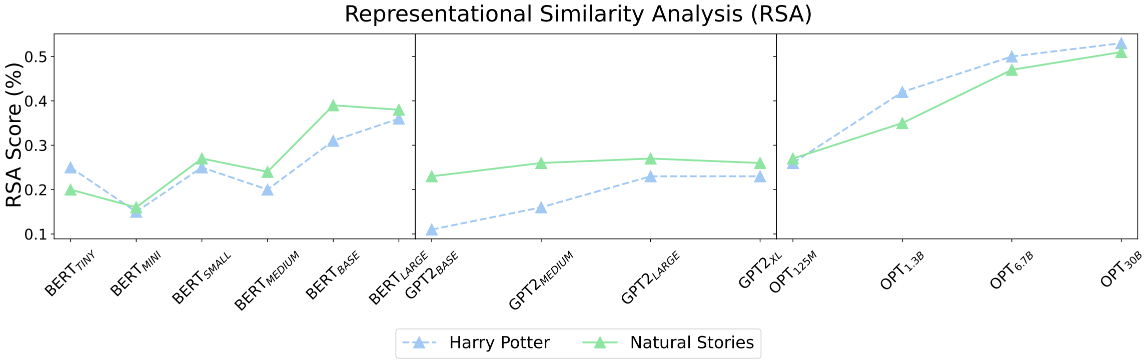

Figure 5: **Convergence results for three families of LLMs using Relational Similarity Analysis**. The correlation score ranges from 0 (no correlation) to 1 (perfect correlation). The plot shows that as the model sizes increase, the representational similarities increase also.

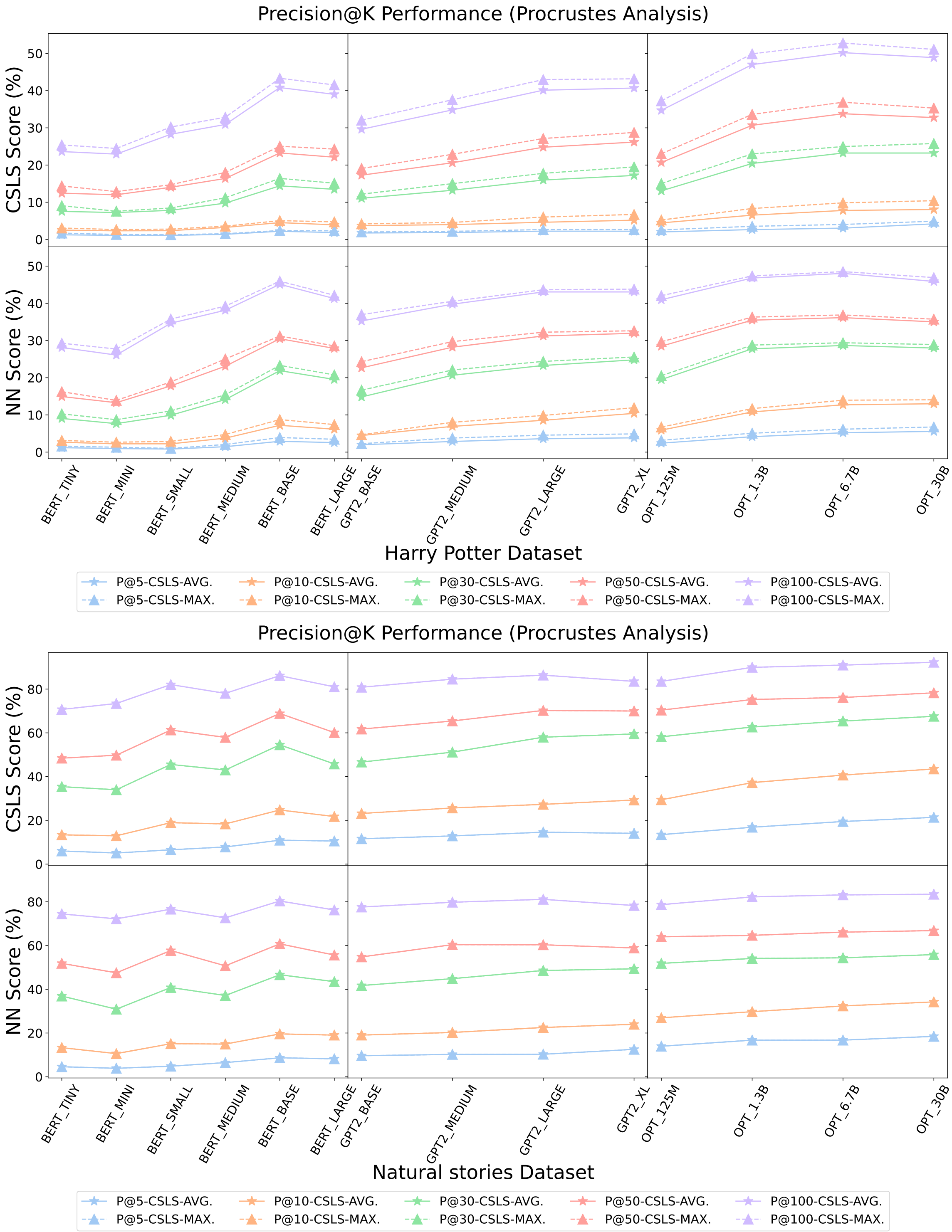

Figure 6: **Convergence results for three families of LLMs across two datasets, using Procrustes Analysis with Gaussian random projection**. The task here is: Given a neural response, which word (in a vocabulary of 1,291 words, was read at the time the response was recorded? Chance P@10 is < 0.01. Convergence is consistent, and some retrieval rates are remarkably high, decoding almost half of the words correctly to a neighborhood of 10 word forms.

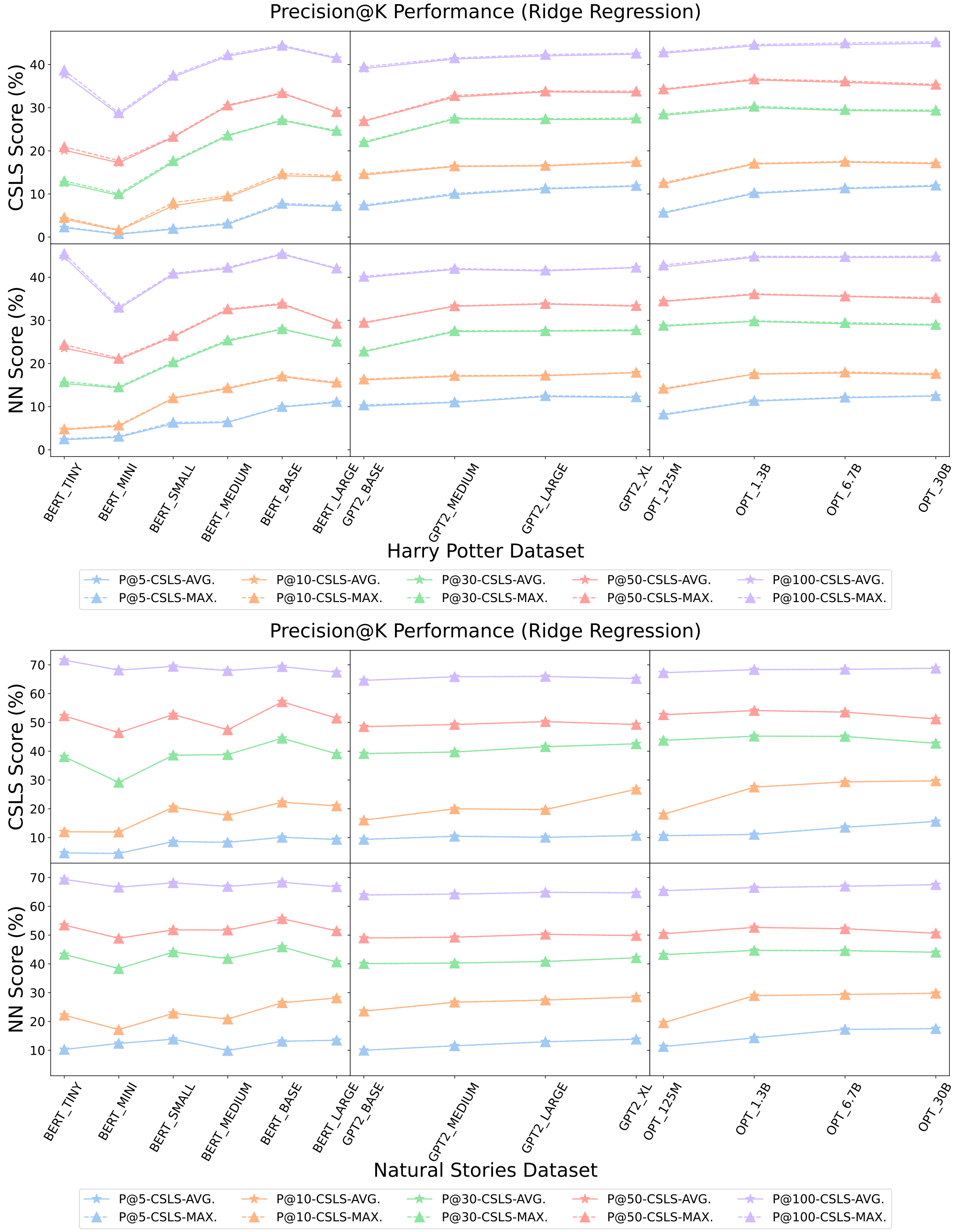

Figure 7: **Convergence results for three families of LLMs using Ridge Regression**. See main paper for results with Procrustes Analysis.

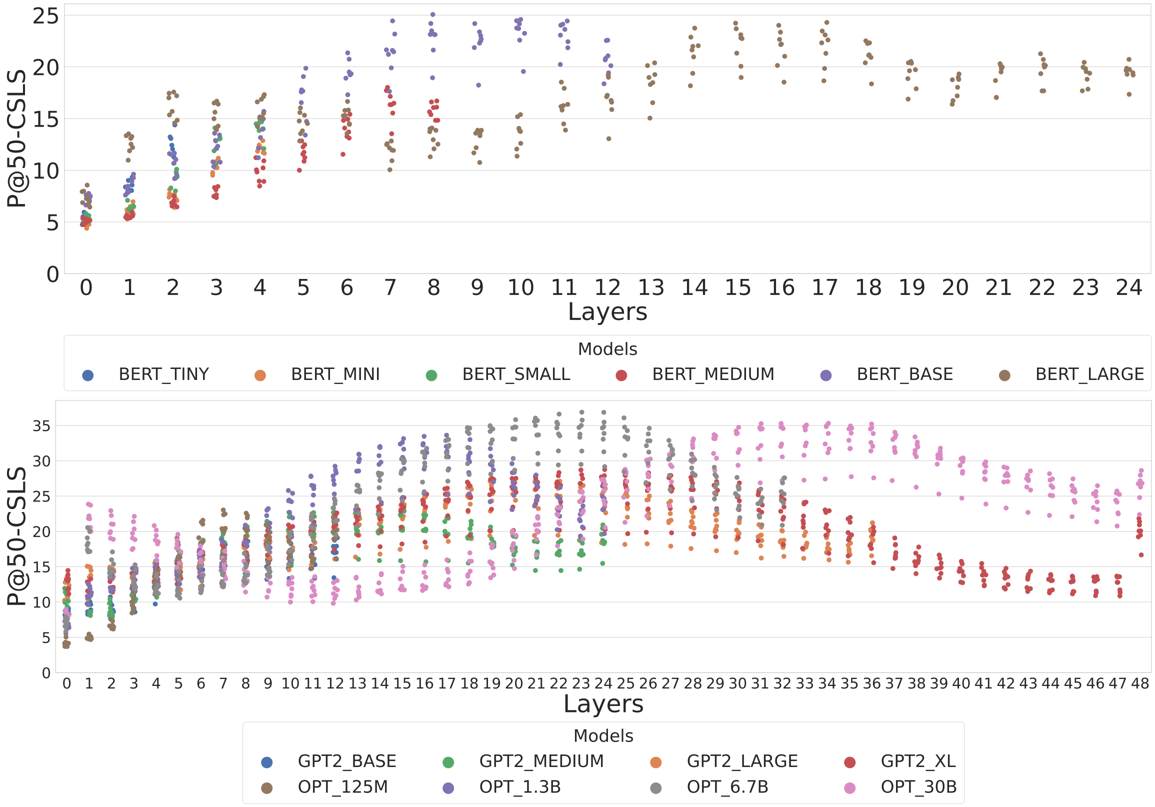

Figure 8: **Alignment precision results across layers.** The plot shows alignment with fMRI (Harry Potter dataset) improves with model depth for LLMs and Procrustes Analysis with Gaussian Random Projection.