# Assessing the alignment between word representations in the brain and large language models

**Elisa Kwon[1], John D. Patterson[2], Roger E. Beaty[2], and Kosa Goucher-Lambert[1]**
[1]*University of California, Berkeley, USA*
[2]*The Pennsylvania State University, USA*
*elisa.kwon@berkeley.edu*

Recent developments in using Large Language Models (LLMs) to predict and align with neural representations of language can be applied to achieving a future vision of design tools that enable detection and reconstruction of designers' mental representations of ideas. Prior work has largely explored this relationship during passive language tasks only, e.g., reading or listening. In this work, the relationship between brain activation data (functional imaging, fMRI) during appropriate and novel word association generation and LLM (Llama-2 7b) word representations is tested using Representational Similarity Analysis (RSA). Findings suggest that LLM word representations align with brain activity captured during novel word association, but not when forming appropriate associates. Association formation is one cognitive process central to design. By demonstrating that brain activity during this task can align with LLM word representations, insights from this work encourage further investigation into this relationship during more complex design ideation processes.



*Design Computing and Cognition DCC'24*. J.S. Gero (ed),
pp. xx-yy. 

## Introduction

Recent advances showcasing the potential of Large Language Models (LLMs) to predict and align with neural representations of language present exciting opportunities for research at the intersection of design, computation, and cognition. Leveraging artificial intelligence (AI) approaches conventionally used to model and generate language, models of language in the human brain have been developed (e.g., [1–3]). While reasons underlying brain-LLM alignment are not fully understood [4], recent work supporting this relationship motivates the present study of brain-LLM alignment during language generation. Broadly, implications for demonstrating this relationship are wide reaching, such as to facilitate the decoding and reconstruction of language from mental representations alone (e.g., by [1]). However, testing the alignment between brain activity and LLMs has mostly been limited to tasks involving passive language reception (e.g., reading text [2] or listening to speech [1,3]). Our work explores this relationship during language generation through word association, a cognitive process engaged during design, especially when forming novel associations [5]. Applied to design, establishing the capabilities of LLMs to semantically model neural representations of language can lead to actualizing a future vision of design tools that effectively utilize brain-machine interfaces. Brain-machine design tools that detect and decode designers' ideas from their minds can enable seamless representation of ideas from the mind and opportunities to provide real-time aid for designers.

In order to realize these brain-machine interactions, in this work, we initially assess whether LLMs produce brain-like responses to language during word generation. Thus, the first research question examined in this work (**RQ1**) is: *Do semantic representations of words produced by LLMs align with neural representations of words during word generation?* This research question is addressed through a word association task, conducted during functional magnetic resonance imaging (fMRI). In this task, subjects (N=35) were instructed to generate either appropriate or novel single word associates to single word stimuli. We assess the alignment between layer-by-layer LLM activations of word prompts to brain responses collected while thinking of associations to the same words. The processes of obtaining and comparing neural and LLM word activations are detailed fully below.

Comparing the two task conditions, we further assess whether differences in brain-LLM alignment exist when thinking of appropriate compared to novel associations of words. This difference is investigated in our second research question (**RQ2**): *Do task goals differently impact brain-LLM alignment during word generation?* As novel association generation is specifically involved in creative thinking, this comparison first contributes

to a fundamental understanding of neurocognitive processes underlying creativity. Through this comparison, we aim to reveal insight into how LLMs may be effectively utilized in design to 'think' creatively in response to a given design prompt. LLM word representations that match well to brain responses during novel word association may also be suitable for directly *generating* creative output. Implications for differences in brain-LLM alignment during appropriate compared to novel word association are explored and discussed in this paper.

## Background

The present work extends upon novel developments in the utilization of artificial intelligence (AI) and LLMs to model language representations by humans in the brain. Key findings from this emerging research area at the intersection of neuroscience and computation are introduced in this section. This work aims to leverage these techniques toward modeling brain-based representations of language during a word association task. Prior work demonstrating the role of association in design and neurocognitive processes underlying design are also reviewed in this section.

### Modeling neural representations of language with language models

With the proliferation of use of artificial intelligence (AI) and neural networks in semantic modeling, significant advancement in encoding and decoding brain-based representations of language has been observed in recent work. The development of encoding models is a data-driven approach that has been used to model voxel-wise brain responses to language [6]. Related works effectively utilizing this approach are presented to motivate this work. Insights from a variety of studies across modalities of semantic stimuli, brain imaging techniques, and language models are introduced.

One frequently used stimulus type administered to elicit language-related brain responses is audio recording (i.e., spoken natural speech). Encoding models built on word features extracted from frequency-based embeddings [3] or word2vec [7] have been shown to predict fMRI BOLD (functional magnetic resonance imaging blood-oxygen-level-dependent) responses in the brain, recorded while listening to natural speech. Défossez et al. reported similar findings when applying a pretrained speech module (wav2vec 2.0) to decode brain responses to speech recorded noninvasively using MEG (magneto-encephalography) and EEG (electro-encephalography) [8]. More recently, Tang et al. demonstrated how a generative neural network language model can be applied to reconstruct continuous language from fMRI activity during natural speech listening [1].

To predict fMRI and MEG responses recorded during sentence reading, Caucheteux et al. trained models with language transformers, and observed convergence between the deep learning algorithms and brain responses [2]. Toneva & Wehbe also used fMRI recordings while reading complex natural text to understand how transformer models (e.g., BERT) encode information relevant to language processing, additionally finding that middle model layers best encode longer sentences [9]. The success of word representations by LLMs to enable brain encoding and decoding motivates our exploration of LLMs in this study.

These examples of prior work, briefly introduced here, have largely explored the relationship between language models and neural representations of language during *reception*, i.e., listening or reading tasks. Related work has also investigated how language models predict brain response during mental simulation of words [10] or semantic comprehension [11]. Our work investigates brain-LLM alignment during language *generation*, specifically when forming associations to words. Importantly, to study processes involved in creativity and design, it is essential to assess the effectiveness of these methods when applied to modeling brain activity during new generation and ideation.

**Applying Representational Similarity Analysis (RSA) to compare representations of language**

While encoding models provide an approach to use language models to directly predict voxel-wise brain responses to language, this work instead employs Representational Similarity Analysis (RSA) to initially investigate alignment between brain activity and LLMs during word generation. RSA is an analytical method developed to characterize activity patterns across voxels in the brain using representational dissimilarity matrices (RDMs) [12]. RDMs are matrices of pairwise comparisons (e.g., dissimilarities) of e.g., brain responses to stimuli, which can provide insight into how the brain represents different information.

Applications of RSA also exist beyond the field of neuroscience, such as to compare representation of language by different LLMs [13,14]. Klabunde et al. conduct RSA to compare representational similarities across various 7 billion (7b) parameter LLMs (i.e., LLMs with 7b model weights) [14], finding that they are not universal across models [13]. Relatedly, Kornblith et al. construct RDMs to compare structural similarities between deep neural networks [15]. In this study, RSA is applied to compare similarities between single words as represented in the brain and by an LLM. The brain response associated with each word reflects either appropriate or novel association generation with the given word and LLM responses are layer-by-layer activations of the given word.

### Cognitive processes underlying creativity and design

Exceeding the complexity of prior work using LLMs to model and predict brain activity during simple language tasks, we are furthermore interested in brain-LLM alignment during the design process. As an initial step, in the present study, insights from a word association task reveal not only how humans and LLMs represent semantic information, but additionally how this relationship varies when humans think of words with a common or creative framing. Association is considered essential to creative thinking [16] and forming novel associations has been shown to contribute to the design process [4]. Even performance on simple tasks such as single-word association demonstrates a relationship with individual creativity measures [17,18]. In a design task, Yin et al. found that high creativity individuals engaged in remote association processes and utilized more association processes than low creativity individuals [5]. Neuroimaging methods, such as EEG [19] or fMRI [20], have been used to identify neural differences during remote compared to common association in creativity tasks.

While design tasks are distinct from creativity tasks, similarities have been observed at a neural level between brain activation during basic creativity tasks and during design studies [21,22]. Investigating ideation processes of product design engineers using fMRI, Hay et al. found alignment in brain activation patterns with reported findings from generic creative ideation tasks [21]. Goucher-Lambert et al. observed that when deriving design ideas with inspirational stimuli, brain activation patterns were consistent with neural correlates to creativity-relevant tasks, such as semantic processing, word representation, and word meaning/retrieval [19]. At a basic level, brain activity observed during creativity can also be reflected in design relevant tasks.

In this paper, we test whether differences in brain activation patterns that emerge during appropriate vs. novel word association contribute to how successfully they match with LLM word representations. This relationship is explored by determining how brain-LLM alignment is impacted by the generation of appropriate compared to novel associates of words.

## Methods

The main aim of this work is to assess the alignment between neural and LLM representations of words. To collect neural representations of words, an fMRI study was conducted in which subjects completed a word association task to generate appropriate and novel word associates to provided stimuli. LLM activations from Llama-2 7b for the same sets of

stimuli were produced for individual layers of the model to compare to these neural representations. Enabling this comparison, Representational Similarity Analysis (RSA) techniques are used, as described in this section.

## Word association task

### *Participants*

A total of 35 young adults (students) participated in the study. Participants received cash payment for their involvement. All participants were right-handed with normal or corrected-to-normal vision and reported no history of neurological disorder. One participant was excluded who failed to complete the task (24 females; mean age: 20; age range: 18–31). The study was approved by the Penn State Institutional Review Board. Informed consent was obtained prior to participation.

### *Study procedure*

A simple single word-association task was conducted in this study, consisting of a generation and evaluation phase. Data from the evaluation phase was not analyzed in the present work; but full details are available (see [23]). In the generation phase, participants were instructed to either generate an appropriate or novel association to a given word stimulus (e.g., noun = 'belt', appropriate association = 'pants', novel association = 'stars'). Participants were asked to generate associations that were concrete nouns. Post-task analyses in prior work revealed higher semantic distances between cue words and associated responses in the novel condition, ensuring that task instructions facilitated differences between the two conditions [23].

Following a 5s pre-instruction fixation, the association instruction appeared and lasted 5s. After a 4–6s jittered fixation cross presentation, a noun from the stimulus list appeared on the screen for 1s. Participants were then given 5s to generate an association, which was immediately followed by a 3s window to orally provide their response. If a participant could not think of an association, they were instructed to say "none." This process is summarized in Fig. 1 for a single block and trial.

### *Task stimuli selection*

Participants completed the word association task in the fMRI scanner where they were presented with a total of 60 nouns during the generation phase (12 trials per run; 5 runs total). Each run of the generation phase consisted of two blocks (six appropriate and six novel trials), listed in Table 1.

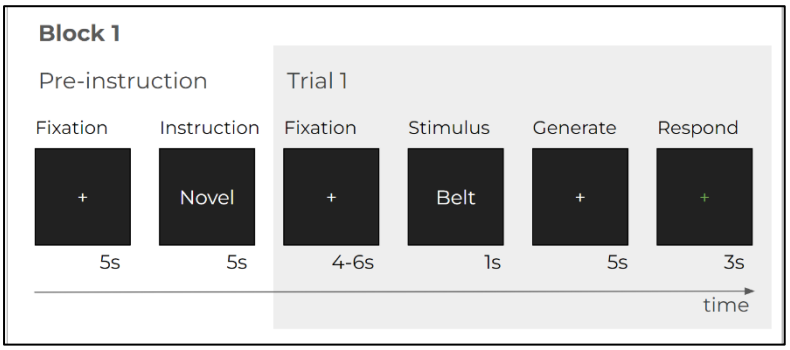


**Fig. 1** Word association task and trial procedure in generation phase of study

**Table 1** Word stimuli presented during the word association task

| | Run number | | | | |
|---|---|---|---|---|---|
| Association type | 1 | 2 | 3 | 4 | 5 |
| Novel | coin<br>hat<br>statue<br>alley<br>subway<br>sword | rock<br>plant<br>palace<br>branch<br>bath<br>plane | seat<br>match<br>log<br>church<br>rocket<br>pet | purse<br>leather<br>ladder<br>robot<br>brass<br>shell | mail<br>dock<br>coat<br>pocket<br>screen<br>hotel |
| Appropriate | bench<br>train<br>steam<br>page<br>sink<br>fence | belt<br>tea<br>net<br>earth<br>tray<br>barn | wheel<br>drill<br>carpet<br>circus<br>costume<br>gym | rain<br>bar<br>engine<br>grass<br>shadow<br>glove | bucket<br>gum<br>sea<br>map<br>pole<br>drum |

The presentation of appropriate or novel stimuli first was counterbalanced across two groups and alternated between runs. Stimuli were selected from a database of 1716 nouns that appeared in several publicly available databases of psycholinguistic norms and reduced using the six following criteria: word frequency, concreteness, imageability, valence, semantic diversity, and cue set size. This yielded a reduced list of 298 words, which was further reduced by manually removing all animate words (humans, animals, professions, body parts), resulting in 160 words. From these, random lists of 30 words were selected (1 list for the novel condition, 1 list for the appropriate condition), until there were no significant differences on any of the six word features (according to t-test analyses; see full analyses and psycholinguistic features of each stimulus list in [23]).

## fMRI data analysis

The fMRI data collected during the task was acquired and preprocessed following the steps outlined in this subsection. In the subsequent analyses performed, only fMRI data in the language-network region of interest [24] was utilized, visualized in Fig. 2.

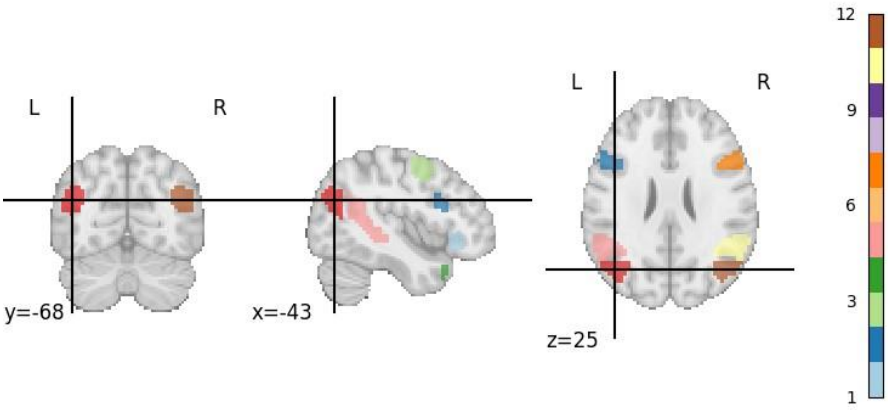


**Fig. 2** Selected regions of interest from [24] are projected on a standard template image (MNI152). A total of 12 parcels are highlighted, including: in each hemisphere, three frontal parcels (inferior frontal gyrus [IFG], its orbital portion [IFGorb], and middle frontal gyrus [MFG]) and three temporal/parietal ones (anterior temporal [AntTemp], posterior temporal [PostTemp], and angular gyrus [AngG]).

**fMRI Data Acquisition:** Structural and functional images were acquired using a Siemans 3 T scanner equipped with a 20-channel head coil. Structural images were acquired with a 2300ms TR, 2.28ms TE, 256 mm field of view (FOV), 192 axial slices, and 1 mm slice thickness. Echo-planar functional images were acquired using an interleaved acquisition, 2500ms TR, 35ms TE, 240mm FOV, 90◦ flip angle, 42 axial slices with 3mm slice thickness resulting in 3mm isotropic voxels.

**Anatomical Data Preprocessing:** T1-weighted (T1w) images were corrected for intensity non-uniformity using N4BiasFieldCorrection (ANTs 2.2.0), which then served as a reference throughout the workflow. Skull-stripping was executed using the antsBrainExtraction.sh workflow (ANTs), with OASIS30ANTs as the target template. Brain tissue segmentation into cerebrospinal fluid (CSF), white-matter (WM), and gray-matter (GM) was conducted on the skull-stripped T1w images utilizing fast (FSL 5.0.9). The spatial normalization to the MNI152NLin2009cAsym standard space was achieved through nonlinear registration of brain-extracted T1w reference and template using antsRegistration (ANTs 2.2.0).

**Functional Data Preprocessing:** Functional data preprocessing was applied to each of the 10 BOLD runs per subject. This involved generating a reference volume and its skull-stripped version using fMRIPrep's custom methodology. BOLD references were aligned to the T1w references using bbregister (FreeSurfer), configured for nine degrees of freedom to address residual distortions. Head-motion parameters were estimated using mcflirt (FSL 5.0.9), followed by slice-time correction of BOLD runs with 3dTshift (AFNI 20160207). The BOLD time-series were then resampled to native space and standard space (MNI152NLin2009cAsym).

### Language model: Llama-2 7b

The LLM used in this study to obtain model-based representations of task stimuli is the open-source, Llama-2 7b-parameter generative text model [25]. The Llama family of models uses a decoder-only transformer architecture (similar to that used in the GPT family) and, relative to open-source models of comparable size, achieves state-of-the-art language understanding and reasoning performance. A key feature of Llama-2 is its emphasis on maintaining a low inference over training budget. In other words, Llama is developed to be fast at inference instead of training, resulting in a smaller model that is trained longer. 4096-dimensional word representations are obtained from Llama-2 for each of 32 hidden layers of the model. These activations are then transformed into a representation that can be quantitatively compared to neural representations using representational similarity analysis, as next described.

### RSA analysis

Representational similarity analysis (RSA) is the analytical technique used in this work to compare how similar activity in the brain is to activity in individual layers of the Llama-2 model. RSA operates by exposing two systems (in this case, a human brain and an LLM) to the same set of conditions (word stimuli). Within each system, distances between representations associated with each pair of conditions are calculated to produce n x n representational dissimilarity matrices (RDMs) that represent the representational geometry of each system under the conditions assessed [12]. Each cell in an RDM reflects the difference in how a pair of conditions/stimuli are represented.

The overall process for developing RDMs from brain and LLM data is summarized in Fig. 3. The RDMs of the two systems can then be compared to assess the degree of representational alignment between them. In the present work, correlation distance (i.e., 1 - correlation) is used to construct RDMs for both the brain and LLM (Fig. 3c-d)—the former based on

voxelwise fMRI activity patterns (Fig. 3a) and the latter based on activity patterns at each hidden layer (i.e., one RDM per layer; Fig. 3b). Pearson correlation, *r*, then compares the fMRI RDM to the LLM RDMs (Fig. 3e). The processes for developing RDMs from the collected fMRI data and from LLM word representations are next outlined.

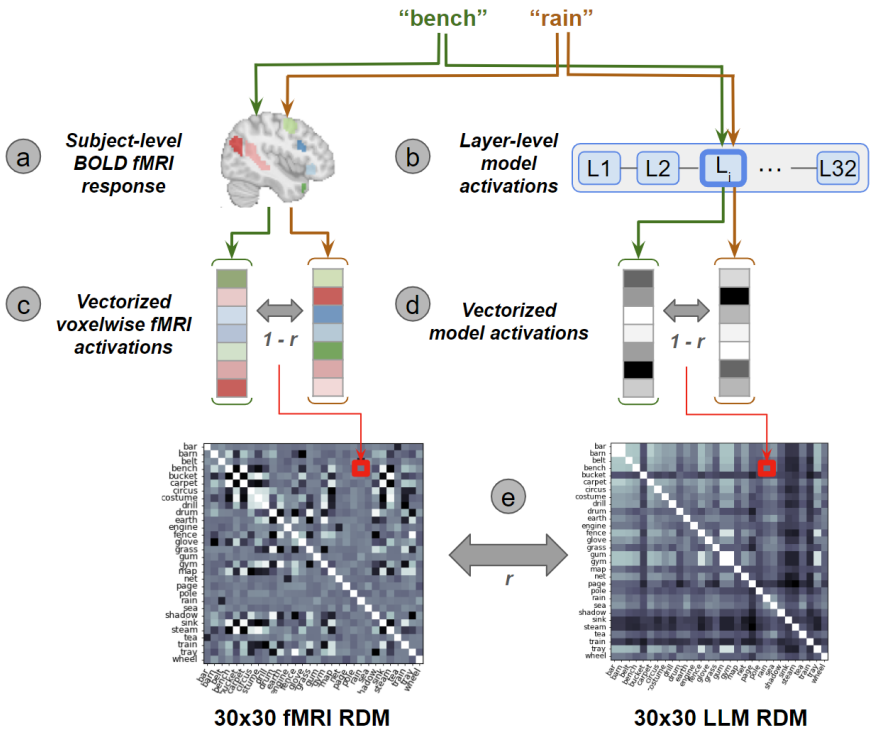


**Fig. 3** Process of developing fMRI and LLM RDMs. For each stimulus pair **a**) subject-level brain response and **b**) layer-level LLM activations are obtained. Dissimilarities between vectorized **c**) fMRI and **d**) LLM-based multidimensional activations for word are computed using 1 – correlation (1-*r*). **e**) fMRI and LLM RDMs are constructed from pairwise dissimilarities and compared for each subject-LLM layer with Pearson correlation, *r*.

### *fMRI RDM construction*

For each subject, two separate 30x30 RDMs were constructed to represent dissimilarities in brain activity patterns between words: the first, for stimuli in the appropriate condition, and the second for words in the novel condition. The preprocessed fMRI time-series data from the targeted language-network region of interest [24] was first used to model the hemodynamic response to stimuli for each participant. A General Linear Model (GLM)

was constructed via the PyMVPA library in Python using stimulus/word and run number as conditions in the model [26]. The model was fit to brain activation data across the entire 5s generation phase (1s during stimulus presentation and 4s during word generation). fMRI RDMs were then constructed by computing pairwise dissimilarities (1-*r*) between fMRI activation patterns modeled by the GLM.

***LLM RDM construction***

Beyond applications in neuroscience, RSA has also been used in prior work to compare language representations of different LLMs [13,15]. In the present work, each word was tokenized and passed as an input to the model; the 4096-dimensional activations corresponding to each word were then extracted from each of Llama-2's hidden layers. To generate LLM RDMs, pairwise dissimilarities (1-*r*) between the model activations for each word were computed using the rsatoolbox library in Python (https://rsatoolbox.readthedocs.io); this was done for each layer in Llama-2. Thus, each cell in the LLM RDMs represents the dissimilarity in the model's output for each pair of words. For each of Llama-2's 32 layers, 30x30 RDMs were produced for both stimuli in the appropriate and novel conditions.

## Results and Discussion

As introduced in the previous section, RSA is used in this work to compare brain and LLM-based word representations. To compare the fMRI and LLM RDMs, Pearson correlations are computed between vectorized upper triangles of the constructed RDMs (symmetric around the diagonal). Addressing **RQ1**, we determine whether there is alignment between brain and LLM-based word representations by computing average fMRI-LLM RDM correlations across participants at each layer, for both task conditions (appropriate and novel word generation). Secondly, to address **RQ2**, fMRI-LLM RDM alignment between brain and model responses to words presented during appropriate and novel word generation are compared.

### Assessing brain-LLM alignment across participants and LLM layers

The first research aim of this study is to understand whether there is alignment between brain and LLM-based word representations. In Fig. 4, the relationships between fMRI and LLM-based RDMs are illustrated across 32 layers of Llama-2 7b. Each bar in Fig. 4 visualizes the average correlation between participants' fMRI RDMs and the LLM RDM constructed to

represent dissimilarities between stimuli in each condition by the specified LLM layer. Positive alignment between RDMs is tested and demonstrated by greater than zero average correlation means across participants. Layers for which the average fMRI-LLM RDM correlations are greater than zero (based on one-tailed one-sample t-tests) include layers 3-4, 7-12, 14, 17-18, 23-26 for novel condition RDMs, also indicated in Fig. 4 (e.g., average $r_{novel}$ = 0.02, t(33)=-2.52, $p$<0.01 at layer 9). Effect sizes at this scale are expected in fMRI analyses and observed in related works (e.g.,[2]).

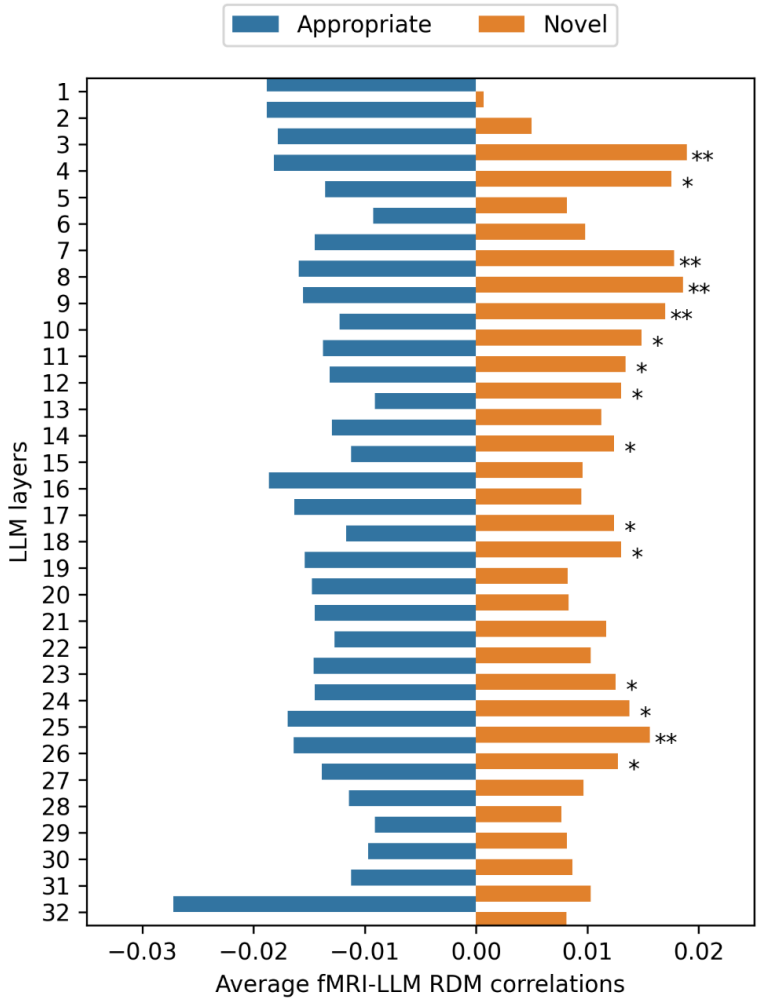


**Fig. 4** Average fMRI-LLM RDM correlations between individual participants' fMRI RDMs and RMDs for each LLM layer. Results for one-sample t-tests against zero (one-tailed, average > 0) shown for each average; * = $p$<0.05, ** = $p$<0.01.

Contrary to observations by Huth et al. [27], asymptotic brain-LLM alignment in later layers of the LLM was not specifically observed in this analysis. Notably, across all layers of the model, average fMRI-LLM RDM correlations for stimuli in the appropriate word generation condition are consistently below zero. The opposite relationship is observed for fMRI-LLM RDM alignment representing stimuli in the novel word generation condition of the task (above zero correlations). These differences are directly assessed and discussed in the following subsection.

## Comparing brain-LLM alignment between task conditions

### *Task condition differences observed in fMRI-LLM RDM correlations*

As shown in Fig. 4, fMRI-LLM RDM alignment appears to differ for fMRI RDMs constructed based on neural representations during novel compared to appropriate word generation. This difference is found to be statistically significant based on paired two-sided two-sample t-tests comparing fMRI-LLM RDM correlations for both conditions at layers 1-12, 14, 16-27, 32 (e.g., average $r_{novel}$ =0.02, average $\underline{r}_{appropriate}$=-0.02, $t_{66}$=-3.64, $p$<0.001 at layer 8). The difference in alignment of fMRI-LLM RDMs between novel and appropriate task conditions is thus observed for activations across most layers of the LLM. Since distinct stimulus sets were used in each task condition, to ensure that these differences in alignment are related to brain activity in each task condition and not stimulus features, additional analyses are performed.

### *Task condition differences observed in word representations in the brain*

Previously, two condition-specific 30x30 fMRI RDMs were developed (following the procedure outlined in Fig. 3a, c). An additional 60x60 fMRI RDM is constructed to assess dissimilarities in brain activity patterns during appropriate vs. novel word generation, as displayed in Fig. 5.

This RDM in Fig. 5 includes dissimilarities in brain activity between forming an appropriate vs. a novel word association, previously missing in the 30x30 fMRI RDMs. By visual inspection of Fig. 5, dissimilarities between brain activation patterns when generating different types of word associations (appropriate vs. novel) appear higher than when generating the same type of word association. This relationship suggests that higher fMRI-LLM RDM alignment during novel word association is related to how words are distinctly represented in the brain during novel vs. appropriate word association (and not related to the stimulus sets seen in the task).

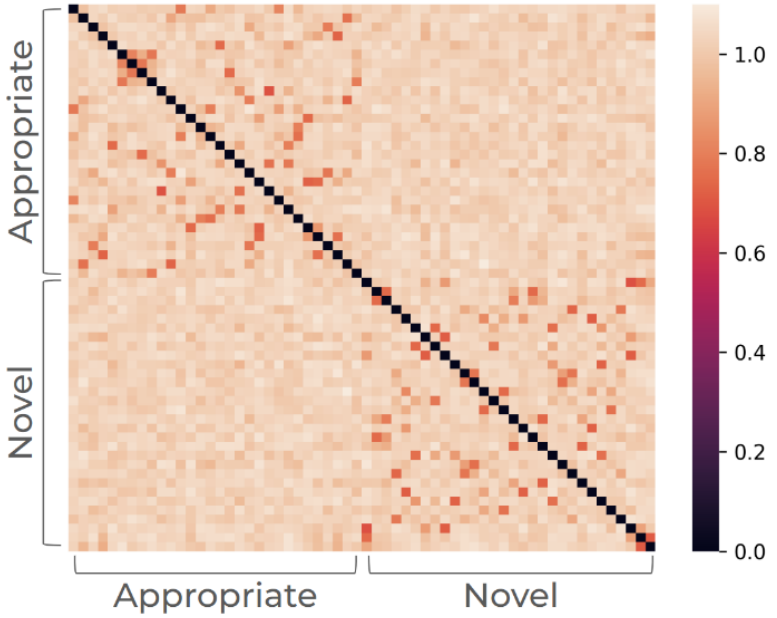


**Fig. 5** RDM (60x60) with pairwise dissimilarities in brain activation patterns of stimuli seen during both appropriate and novel word generation

Average dissimilarities in brain response when performing the same vs. different types of word association are shown in Fig. 6. Using two-tailed two-sample t-tests, the average pairwise dissimilarities in brain response related to appropriate-appropriate, novel- novel, and appropriate-novel word associations are compared.

If average dissimilarities between neural word representations of appropriate-appropriate stimulus pairs and novel-novel stimulus pairs differ, this would indicate a potential effect other than task condition on the previous findings. Instead, we find that when generating the same type of association for two words (i.e., appropriate-appropriate vs. novel-novel), no statistically significant difference in word dissimilarities is observed ($t_{29578}$=0.21, $p$=0.84), whether the associations made are appropriate or novel. However, there are differences observed when comparing dissimilarities between word generation of the same and different types ($t_{29578}$=-19.6, $p$<0.0001, $t_{29578}$=-19.1, $p$<0.0001). In other words, the way the brain represents semantic information appears to differ during novel association compared to appropriate association. This analysis supports our initial result that higher fMRI-LLM RDM alignment for stimuli seen during novel word association is related to differences in brain activity and not stimuli or LLM representations specific to the task condition.

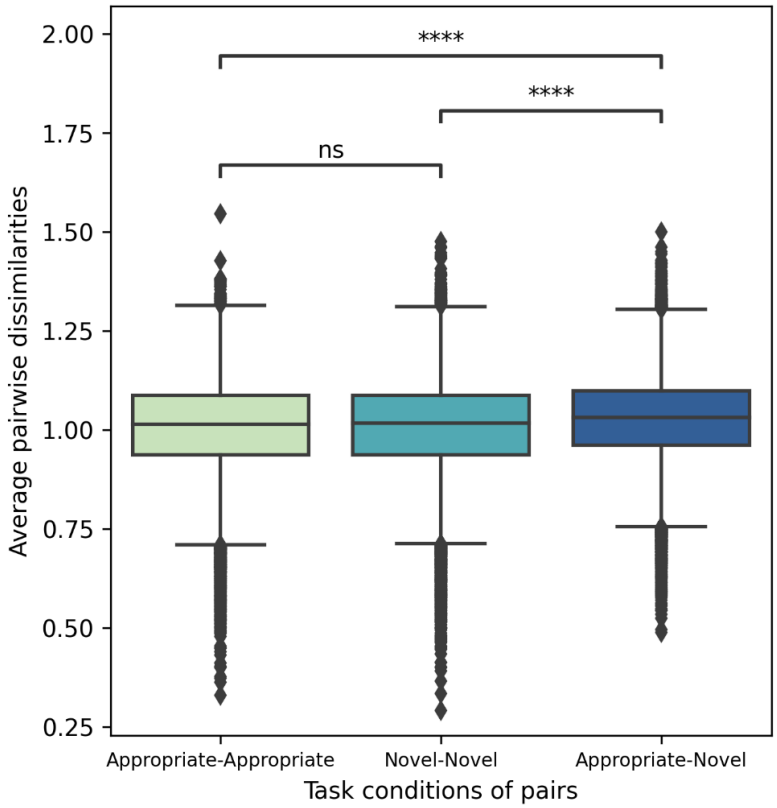


**Fig. 6** Average dissimilarities in brain activation patterns during word generation across different task conditions. Results for paired two-sample t-tests (two-tailed) shown for each average; * = $p$<0.05, ** = $p$<0.01, *** = $p$<0.001, **** = $p$<0.0001.

Based on these findings, we suggest that for LLMs to align with representations of brain activity during active word generation, it may be necessary to engage in deep thinking processes, likely more prominent during novel than appropriate word association. This interpretation is supported by prior work by Soto et al. who observed higher accuracy of LLM-based models to decode brain activity during deep processing (mental simulation of an item's features) than shallow processing (reading and repeating name of item) [10].

## Implications for design

In this study, the relationship between language representations by LLMs and in the brain during language generation was explored. We observed positive alignment in word representations by LLMs and brain activity during language generation, specifically during novel word association. The

alignment of LLMs with human neural representations of semantic information observed during language generation reveals potential for eventual uses of AI and language models toward decoding mental representations. This work contributes encouraging findings towards this vision, and other insights into and applications for design.

Our results support prior work establishing neural correlates of association in design [19] by establishing differences in brain activity patterns representing semantic information during appropriate vs. novel word association. Furthermore, we show that there is higher brain-LLM alignment when generating novel word associates. Demonstrating that LLMs and brain representations can similarly represent semantic information during language generation provides positive insight for applying LLMs to decoding designers' new ideas. Enabling the direct decoding of newly generated words from designers' mental representations can help remove barriers in traditional design contexts that may limit expression and representation of design ideas. Real-time brain decoding also has wide implications for novel brain-machine interfaces for design. The impact of neurocognitive feedback on design ideation output (e.g., cognitive activation feedback [28]) may be enhanced and improved with added context about what designers are thinking. Prior work has shown how providing designers with inspirational stimuli selected at varied analogical distances from their idea in real time can affect different design outcomes [29]. Applied in a brain-machine interface, a designer could be provided with inspirational aid based on their design idea as it emerges, supporting their continued ideation and design process.

In creativity and design research, some applications of LLMs thus far have been to generate design concepts [30] or score creativity of human-generated responses using semantic distance measures [31]. Different from these applications, our work proposes that LLMs can directly decode human-generated design concepts based on their mental representations alone. The findings observed in this study provide encouraging directions for further exploration into modeling brain responses with LLMs toward new design tools and applications.

## Limitations and future work

In this paper, RSA is used to compare RDMs constructed to illustrate word representations by LLMs with word representations in the brain during active association generation. Potential limitations of this study and opportunities for future work are discussed to advance these findings.

This study explores differences in brain-LLM alignment during appropriate and novel word association generation across multiple levels of granularity, with the most granular related to group averages across participants. As this relationship may vary across individuals, future work can further explore how individual differences may impact these findings.

A full comparison of RDMs consisting of dissimilarities between all 60 words in the study was not performed, due to the difference in stimuli used in each condition of the task. Although efforts were made to match psycholinguistic features of words in the appropriate and novel stimulus sets (reported in [23]), stimulus features may contribute to observed differences between brain activity in these conditions. While a limitation, our findings suggest that the differences observed are related to the task condition engaged, rather than to features of words in the stimulus sets presented.

The neural representations of words studied in this paper were obtained in an MRI scanner, where participants were lying supine viewing the projected text. The use of fMRI in design research must consider trade-offs between fMRI constraints (e.g., noisy, use in an enclosed space) and ecological validity [32]. In this work, a simple task was engaged to initially examine brain activity during language generation. For continued study on brain-LLM alignment in design, brain activity during design concept generation may be desired, for which fMRI usage may be limited. Future work should explore adapting experimental paradigms for design research to be suitable for study using fMRI, or ways of utilizing brain signal obtained from more portable devices e.g., EEG or fNIRS (functional Near Infrared Spectroscopy) towards brain decoding (e.g., [33]).

Related to the LLM word representations obtained, in this initial investigation, we examined decontextualized LLM word representations only. However, humans completed the task under performance-guiding context (instructions) that may have altered neural responses relevant to generating associates. Thus, an important next step for this research is to investigate whether and how the provision of task context to the LLM affects alignment with human brain data. Related to model selection, although the 7b model (the smallest Llama-2 model) has been shown to be performant relative to other models of comparable size on NLP benchmarks, the 70b version performs approximately twice as good [25]. Antonello et al. for example, show that bigger models tend to better match human brain data during a language reception task [27]. As such, examining the influence of model size on brain-machine alignment is another promising avenue for future work. As Klabunde et al. observed [13], representational similarity is not universal across different models. This work therefore invites further investigation of the alignment between different language models and brain activity.

Pearson correlation is used as the main metric of correlation and dissimilarity in this study. Other measures of computing similarity when constructing and comparing RDMs are available, and different approaches to assessing alignment can also be explored. Decoded LLM activations at intermediate states, for example, can be compared with participants' generated word associations. This analysis can contribute to providing insight to results specific to alignment with individual model layers.

Finally, beyond language, computational models that represent semantic and visual features of images (e.g., ResNet50) have been effectively leveraged to decode brain response when viewing images [34]. Investigating this relationship with the future aim of decoding mental representations of new ideas is another exciting direction for future exploration.

## Conclusion

An overarching objective of this work is to investigate whether LLMs can be suitable candidates for applications in decoding mental representations of generated ideas or design concepts. Prior work has thus far not widely tested the relationship between language representation by LLMs and brain activity in humans during language generation tasks. Our findings suggest that not only do neural activations during word generation and LLM-based representations of words align, but this relationship is demonstrated during the formation of novel compared to appropriate word associations specifically. While somewhat unintuitive, since LLMs may be expected to 'think' of words in terms of natural, appropriate, or expected associations, this relationship suggests that the deeper mental processing required to form novel word associates promotes improved alignment. This finding encourages further investigation into brain-LLM alignment during complex tasks including, beyond generating simple novel word associations, other cognitive processes involved in creativity and design.

## References

1. Tang J, LeBel A, Jain S, Huth AG (2023) Semantic reconstruction of continuous language from non-invasive brain recordings. Nat Neurosci 26(5):858–66.
2. Caucheteux C, King JR (2022) Brains and algorithms partially converge in natural language processing. Commun Biol 5(1):1–10.
3. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532(7600):453–8.

4. Oota S, Gupta M, Toneva M (2023) Joint processing of linguistic properties in brains and language models. Advances in Neural Information Processing Systems. 36:18001–14.
5. Yin Y, Zuo H, Childs P (2023) Impacts of cognitive factors on creativity quality in design: identification from performances in recall, association and combination. Journal of Intelligence 11(2):39.
6. LeBel A, Wagner L, Jain S, Adhikari-Desai A, Gupta B, Morgenthal A, et al. (2023) A natural language fMRI dataset for voxelwise encoding models. Sci Data 10(1):555.
7. Zhang Y, Han K, Worth R, Liu Z (2020) Connecting concepts in the brain by mapping cortical representations of semantic relations. Nat Commun. 11(1):1877.
8. Défossez A, Caucheteux C, Rapin J, Kabeli O, King JR (2023) Decoding speech perception from non-invasive brain recordings. Nat Mach Intell. 5(10):1097–107.
9. Toneva M, Wehbe L (2019) Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems p. 14928–38.
10. Soto D, Sheikh UA, Mei N, Santana R (2020) Decoding and encoding models reveal the role of mental simulation in the brain representation of meaning. Royal Society Open Science 7(5):192043.
11. Caucheteux C, Gramfort A, King JR (2022) Deep language algorithms predict semantic comprehension from brain activity. Sci Rep. 12(1):16327.
12. Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2.
13. Klabunde M, Amor MB, Granitzer M, Lemmerich F (2023) Towards measuring representational similarity of large language models. arXiv; Available from: http://arxiv.org/abs/2312.02730
14. Klabunde M, Schumacher T, Strohmaier M, Lemmerich F (2023) Similarity of neural network models: a survey of functional and representational measures. arXiv; Available from: http://arxiv.org/abs/2305.06329
15. Kornblith S, Norouzi M, Lee H, Hinton G (2019) Similarity of neural network representations revisited. arXiv; Available from: http://arxiv.org/abs/1905.00414
16. Beaty RE, Kenett YN (2023) Associative thinking at the core of creativity. Trends in Cognitive Sciences 27(7):671–83.
17. Olson JA, Nahas J, Chmoulevitch D, Cropper SJ, Webb ME (2021) Naming unrelated words predicts creativity. Proceedings of the National Academy of Sciences of the United States of America 118(25).
18. Prabhakaran R, Green AE, Gray JR (2014) Thin slices of creativity: Using single-word utterances to assess creative cognition. Behav Res. 46(3):641–59.
19. Yin Y, Wang P, Han J, Zuo H, Childs P (2023) Comparing designers' EEG activity characteristics for common association and remote association. In: Gero JS, editor. Design Computing and Cognition'22. Cham: Springer International Publishing. p. 255–67

20. Benedek M, Jurisch J, Koschutnig K, Fink A, Beaty RE (2020) Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes. NeuroImage 210:116586.
21. Hay L, Duffy AHB, Gilbert SJ, Lyall L, Campbell G, Coyle D, et al. (2019) The neural correlates of ideation in product design engineering practitioners. Design Science 5:e29.
22. Goucher-Lambert K, Moss J, Cagan J (2019) A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli. Design Studies 60:1–38.
23. Matheson HE, Kenett YN, Gerver C, Beaty RE (2023) Representing creative thought: A representational similarity analysis of creative idea generation and evaluation. Neuropsychologia 187:108587.
24. Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Kean H, et al. (2022) Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. Sci Data. 9(1):529.
25. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models [Internet]. arXiv; Available from: http://arxiv.org/abs/2307.09288
26. Hanke M, Halchenko Y, Sederberg P, Olivetti E, Fründ I, Rieger J, et al. (2019) PyMVPA: a unifying approach to the analysis of neuroscientific data. Frontiers in Neuroinformatics 3.
27. Antonello R, Vaidya A, Huth AG (2023) Scaling laws for language encoding models in fMRI. arXiv; Available from: http://arxiv.org/abs/2305.11863
28. Hu M, Shealy T, Milovanovic J, Gero J (2022) Neurocognitive feedback: a prospective approach to sustain idea generation during design brainstorming. International Journal of Design Creativity and Innovation 10(1):31–50.
29. Goucher-Lambert K, Gyory JT, Kotovsky K, Cagan J (202) Adaptive inspirational design stimuli: Using design output to computationally search for stimuli that impact concept generation. Journal of Mechanical Design 142(091401).
30. Ma K, Grandi D, McComb C, Goucher-Lambert K (2023) Conceptual design generation using large language models. Proc. of ASME 2023 IDETC/CIE (DTM), Boston, MA, Aug 20-23. p. V006T06A021.
31. Beaty RE, Johnson DR (202) Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behav Res. 53(2):757–80.
32. Hay L, Duffy AHB, Gilbert SJ, Grealy MA (2022) Functional magnetic resonance imaging (fMRI) in design studies: Methodological considerations, challenges, and recommendations. Design Studies 78:101078.
33. Wang P, Peng D, Yu S, Wu C, Wang X, Childs P, et al. (2022) Verifying design through generative visualization of neural activity. In: Gero JS, editor. Design Computing and Cognition'20. Cham: Springer International Publishing. p. 555–73.
34. Wang AY, Kay K, Naselaris T, Tarr MJ, Wehbe L (2023) Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nat Mach Intell 5(12):1415–26.