

Large Language Models Align with the Human Brain during Creative Thinking

Mete Ismayilzada^{1,2} Simone A. Luchini^{4,5} Abdulkadir Gokce¹ Badr AlKhamissi¹
Antoine Bosselut¹ Antonio Laverghetta Jr.³ Lonneke van der Plas^{†2}
Roger E. Beaty^{†5}

¹EPFL, ²Università della Svizzera italiana (USI), ³Wesleyan University,

⁴Paris Brain Institute (ICM), ⁵Pennsylvania State University

Abstract

Creative thinking is a fundamental aspect of human cognition, and divergent thinking—the capacity to generate novel and varied ideas—is widely regarded as its core generative engine. Large language models (LLMs) have recently demonstrated impressive performance on divergent thinking tests and prior work has shown that models with higher task performance tend to be more aligned to human brain activity. However, existing brain-LLM alignment studies have focused on passive, non-creative tasks. Here, we explore brain alignment during creative thinking using fMRI data from 170 participants performing the Alternate Uses Task (AUT). We extract representations from LLMs varying in size (270M–72B) and measure alignment to brain responses via Representational Similarity Analysis (RSA), targeting the creativity-related default mode and frontoparietal networks. We find that brain-LLM alignment is positively associated with model size (default mode network only) and with idea originality (both networks), with these relationships clearest early in the creative process. We further find that post-training objectives are associated with functionally selective differences in alignment: a creativity-optimized LLAMA-3.1-8B-INSTRUCT retains alignment with high-creativity neural responses while lacking the positive low-creativity alignment present in other variants, but a reasoning-trained variant shows the opposite asymmetry, with negative alignment to high-creativity responses. Together, these results suggest that brain alignment offers an informative lens on how post-training relates to the neural geometry of human creative thought.

1 Introduction

Creative thinking stands as one of the most distinctive and complex capacities of the human mind, underpinning scientific discovery, artistic expression, and everyday problem-solving (Torrance, 1974; Guilford, 1967; Duncker & Lees, 1948). Within the broader landscape of human cognition, creativity is commonly decomposed into two complementary modes: *convergent thinking*, which involves narrowing down multiple possibilities to arrive at a single, well-defined solution, and *divergent thinking*, which involves generating a wide range of novel ideas or associations from a single starting point (Guilford, 1967; Mednick, 1962). While both modes contribute to creative cognition, divergent thinking is widely considered the more foundational driver of creativity, as it captures the generative, open-ended exploration that gives rise to original thought (Runco & Acar, 2012). Its study has therefore attracted greater attention in both cognitive psychology and neuroscience, with researchers seeking to understand not only what makes humans capable of such flexible ideation, but also what neural substrates and cognitive mechanisms support it (Beaty et al., 2018; Abraham, 2013; Beaty et al., 2016).

[†]Equal supervision

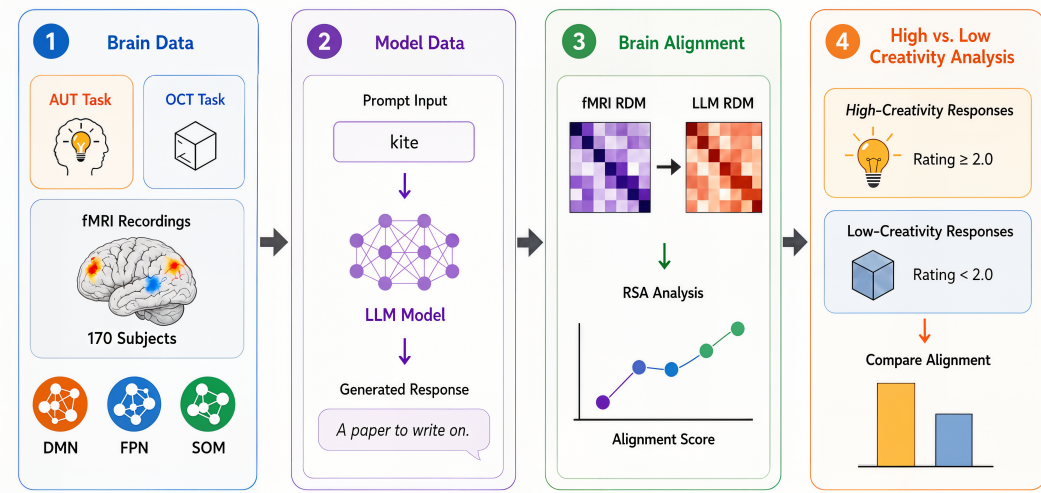


Figure 1: **Our high-level brain alignment methodology.** Participants perform a creative task (AUT) and a matched non-creative control (OCT) during fMRI, while LLMs receive the same instructions and stimuli. We extract layer-wise activations at the prompt stage and after response generation, and compute alignment as the rank correlation between per-subject fMRI and LLM representational dissimilarity matrices (RDMs). Alignment is evaluated in the default mode (DMN) and frontoparietal (FPN) networks, plus the somatomotor network (SOM) as a control, and then compared between high- and low-creativity human responses.

In recent years, large language models (LLMs) have rapidly expanded their reach across a wide range of cognitive tasks, with divergent thinking among the abilities they now exhibit (Ismayilzada et al., 2024a; Zhao et al., 2023). Models have shown impressive results on established creativity benchmarks such as the Alternative Uses Test (AUT), which measures the ability to generate diverse and unconventional uses for everyday objects, in some cases reaching or exceeding average human performance. (Góes et al., 2023; Stevenson et al., 2022; Bellemare-Pepin et al., 2026). Alongside these behavioral results, a growing body of work has established that LLMs with higher task performance tend to exhibit greater alignment with the human language system (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux & King, 2022). However, prior work has predominantly focused on passive language processing tasks, in which participants read or listen to naturalistic text (Aw et al., 2023; AlKhamissi et al., 2025; LeBel et al., 2023) or, recently, simple word association (Kwon et al., 2024) and abstract reasoning tasks (Pinier et al., 2025). Divergent thinking, however, is a higher-order cognitive function often employing a broad and distributed network of brain regions well beyond the classical language system (Liu et al., 2024). Whether more creative LLMs are correspondingly more aligned to the neural representations underlying human creative cognition, therefore, remains an open and compelling question, and one that existing brain-LLM alignment frameworks have yet to address.

In this work, we take a first step toward addressing this question by systematically investigating brain-LLM alignment in the context of creative thinking using the AUT task. We leverage fMRI data collected from Beaty et al. (2018), in which participants performed both a creative task, namely the AUT, and a matched non-creative control task, namely the Object Characteristics Task (OCT), which involves generating typical characteristics of everyday objects (Beaty et al., 2018). Similarly, we extract model representations to the same stimuli in these tasks from a range of LLMs and measure their alignment to the corresponding brain responses using Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). On the neural side, we examine alignment across several canonical regions of interest involved in creative thinking, namely, the Default Mode Network (DMN) and the Frontoparietal Network (FPN) (Luchini et al., 2025; Liu et al., 2024). Our results reveal that brain alignment during creative thinking is stage-dependent: alignment is positively associated with model size and creative task performance when measured at the prompt stage, but this relationship

weakens once models generate responses. We further find that post-training objectives are associated with functionally selective differences in alignment, with creativity-optimized and reasoning-chain training exhibiting opposite asymmetries relative to the neural geometry of human creative thought. To the best of our knowledge, this is the first study to directly examine brain-LLM alignment in the context of an active creative thinking task, and we hope our findings encourage deeper investigation into the neural plausibility of creative cognition in language models.

2 Related Work

Brain-LLM Alignment A growing body of work has established that the internal representations of large language models (LLMs) exhibit meaningful similarity to neural activity recorded during vision (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2016; Gokce & Schrimpf, 2024), auditory (Koumura et al., 2023; Kell et al., 2018; Tuckute et al., 2023), and human language processing (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux & King, 2022; Tuckute et al., 2024; Binz et al., 2025) as well as during comprehension of computer programs (Srikant et al., 2022). These studies have demonstrated that LLM representations can linearly predict brain responses to the same linguistic stimuli across a variety of neuroimaging paradigms, including reading naturalistic stories and sentences. More recent work has extended these findings by examining how specific model properties relate to brain alignment, finding strong correlations with model size and world knowledge, and demonstrating that instruction-tuning further improves alignment beyond what is achieved by pretraining alone (Aw et al., 2023). Parallel efforts have examined how brain alignment evolves over the course of LLM training (AlKhamissi et al., 2025). However, it is notable that prior alignment work has been conducted largely using passive language paradigms in which participants read or listen to naturalistic text. Only recently, some works have examined the brain alignment in active cognitive reasoning tasks such as novel word association generation (Kwon et al., 2024) and abstract reasoning (Pinier et al., 2025). Closest to our work is Kwon et al. (2024), which has evaluated alignment across two task conditions: A creative condition where participants had to generate original associates to a target word, and a non-creative condition where participants would generate appropriate associations. Findings have revealed that LLM activations align with human brain activity during creative word associations, but not during appropriate word associations, indicating that complex language production is crucial to alignment. However, to the best of our knowledge, no prior work has examined brain alignment during active creative thinking. We address this gap by studying brain-LLM alignment in an active creative task (AUT task), extending emerging efforts that probe alignment in settings requiring cognitive flexibility.

LLMs and Creative Thinking Beyond passive language comprehension, LLMs have recently been shown to exhibit impressive performance on tasks requiring creative and divergent thinking (Ismayilzada et al., 2024a). Studies evaluating models on creativity tasks such as the AUT have found that frontier models can generate responses that are rated as highly original, sometimes matching or exceeding average human performance (Góes et al., 2023; Stevenson et al., 2022; Bellemare-Pepin et al., 2026). Prior work has proposed various techniques to improve creative thinking and problem-solving in LLMs, such as creative prompting (Nguyen & Singla, 2025; Tian et al., 2024; Lu et al., 2025) and fine-tuning/preference optimization strategies (Puerto et al., 2025; Ismayilzada et al., 2025). Overall, progress in this field has mainly focused on evaluating and improving divergent creative thinking in LLMs; however, the question of whether high performance on creative thinking tasks is accompanied by representational similarity to the neural systems that support human creativity remains unexplored. We aim to investigate this question of brain-LLM alignment in the present work.

3 Methodology

3.1 Brain Data

We use the fMRI recordings of human subjects while they solve AUT and OCT tasks from [Beaty et al. \(2018\)](#). This brain dataset contains fMRI recordings from 170 healthy subjects. Subjects were given the following instructions, respectively, for each task, and then were shown a stimulus one at a time.

AUT Instructions

Think of creative uses for objects. Report the most original idea. Keep thinking of ideas during the thinking period. Quality is more important than quantity.

OCT Instructions

Think of the physical properties of objects. Report the most prominent characteristic. Keep thinking of ideas during the thinking period.

After a brief period of thinking, their responses, along with fMRI recordings, were collected for each stimulus. In total, subjects solved both tasks for 46 stimuli, where each subject was exposed to on average 23 stimuli for each task. For example, given the stimulus “kite”, one subject responded with “a paper to write on” and “rectangular” respectively for AUT and OCT tasks.

We first perform data cleaning based on the subject responses and, as a result, remove data for 8 participants: those who failed to follow the task instructions, those who failed to provide a response, and those whose preprocessed brain data had a different voxel dimensionality from the remaining subjects, which prevented their inclusion in the group-level RDM analyses. Next, we apply standard fMRI data preprocessing (e.g., removing confounders, detrending, standardizing, and signal filtering) using `nilearn`¹ library, which results in “clean” (but raw) fMRI brain activations ([Matthews & Jezzard, 2004](#)). Rather than using these raw activations, we follow standard practice and estimate single-trial beta maps using a General Linear Model (GLM) ([Draper & Smith, 1998](#)), and conduct all subsequent brain-LLM alignment analyses on these estimated neural response patterns. We extract brain data from across three functional brain networks using the [Yeo et al. \(2011\)](#) parcellation: the Default Mode Network (DMN), the Frontoparietal Network (FPN), and the Somatomotor Network (SOM). The DMN and FPN were selected given their well-established roles in creative cognition ([Beaty et al., 2016](#); [Luchini et al., 2025](#); [Liu et al., 2024](#)), while the SOM was selected given its limited involvement in higher-order cognitive processes, making it a suitable control network.

3.2 Model Data

We provide the same instructions² and the stimuli to an LLM and extract its intermediate layer activations as the model representations of the stimuli. Matching the model’s prompt to the instruction given to human participants is important in brain-LLM alignment studies, since alignment is computed against the neural responses elicited by that instruction; we therefore hold prompt content fixed across conditions in our primary analyses. Task-dependency is instead tested through the OCT control, which matches the AUT in stimulus, structure, and language, differing only in whether the goal is creative generation or object-property reporting ([Beaty et al., 2018](#)). We note that the OCT is a more stringent control than a non-language placeholder prompt would be, as it shares substantially more surface structure with the AUT (42% token overlap). Prior brain-LLM alignment studies typically extract model representations for the input prompt alone, which is sufficient for passive language tasks where no generation occurs. However, this is not sufficient for our task, which involves active generation, so we extract representations at both the prompt stage

¹<https://nilearn.github.io>

²We only replace the word “objects” with the stimulus name since we evaluate language-only models.

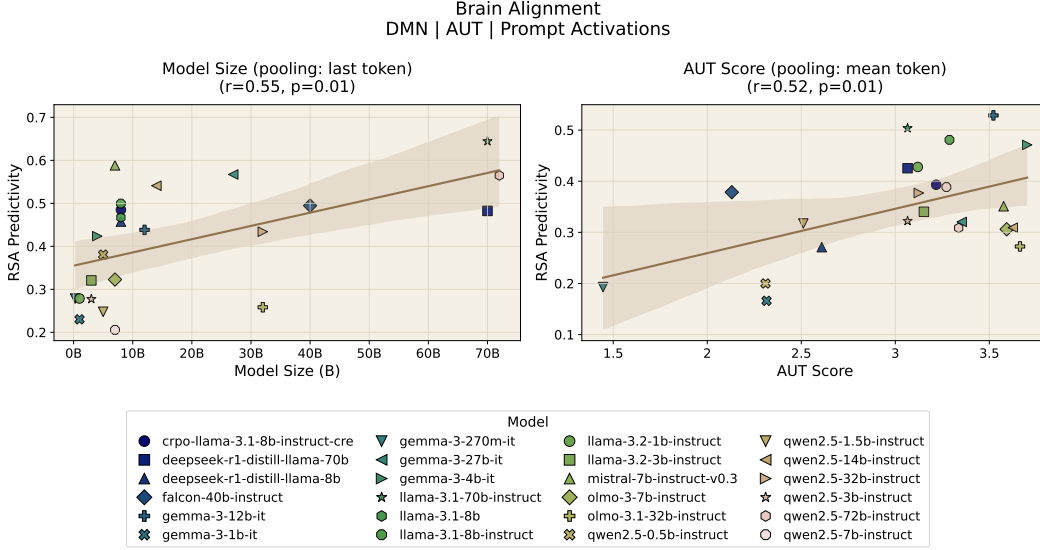


Figure 2: **Default Mode Network (DMN) AUT brain alignment results by model size and task performance using prompt-only model activations.** r and p correspond to the Pearson correlation coefficient and p-value.

and after the model has generated a response, allowing us to capture the full generative process that more closely mirrors the demands of the AUT task. We note that this setup is similar to [Kwon et al. \(2024\)](#) and [Oota et al. \(2025\)](#), except that we examine alignment at both the prompt and the generation stage, while they consider the generation stage only.

3.3 Brain Alignment

Brain alignment refers to the degree of similarity between the internal representations of an LLM and human neural activity ([Schrimpf et al., 2020](#)). We evaluate alignment using Representational Similarity Analysis (RSA) ([Kriegeskorte et al., 2008](#)) following the pipeline conventions of [Nili et al. \(2014\)](#). We note that while some past works have used linear predictivity (i.e., fitting a regression model to predict voxel-wise fMRI responses) ([Schrimpf et al., 2020; 2021](#)), this metric typically requires a large sample of stimuli responses to be effective. While our data is large in terms of the number of subjects, the number of stimulus samples per subject is insufficient to train a meaningful regression model. Therefore, we opt not to use it in our alignment computation (full details in Appendix B).

We compute alignment at the level of individual subjects and report the median across subjects. To account for the inherent noise in fMRI measurements, we estimate a noise ceiling per subject, which represents the maximum alignment score achievable given the reliability of the neural data. Raw alignment scores are then divided by their corresponding noise ceilings, yielding noise-ceiling-normalized scores that allow for more meaningful comparisons across brain regions and models. We evaluate alignment at every layer of each LLM and take the maximum value across layers as the model’s overall brain alignment score, following [Schrimpf et al. \(2020\)](#). This layer-selection procedure is motivated by the observation that different layers of a network capture different levels of representational abstraction and allow each model to be assessed at its most brain-aligned processing stage ([Elhage et al., 2021; Tenney et al., 2019; Jawahar et al., 2019](#)).

3.4 Post-training Objectives and Creativity

Prior work has shown that instruction tuning improves LLM alignment with the human language system ([Aw et al., 2023](#)), raising the question of whether different post-training approaches, such as those targeting creativity or reasoning, can selectively modulate alignment

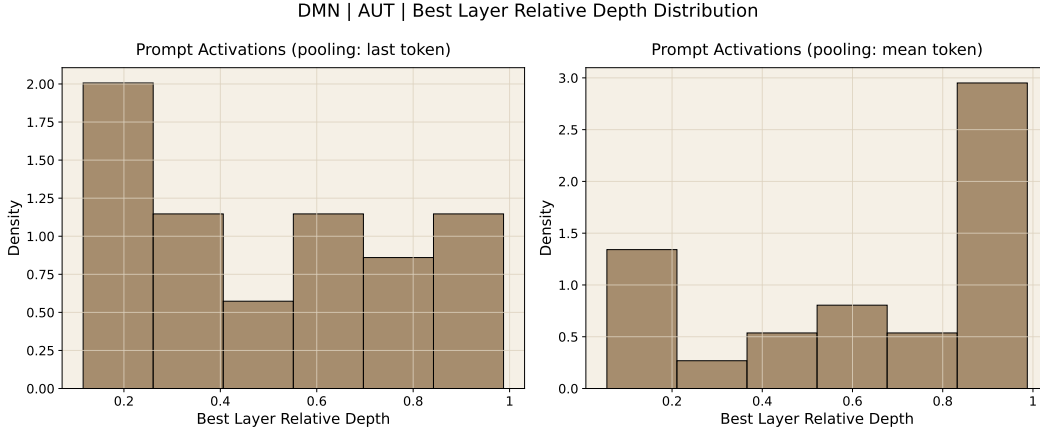


Figure 3: The distribution of best model layer (measured by relative depth) for DMN AUT brain alignment on prompt-only model activations.

with neural responses associated with creative ideation. To explore this, we partition the human brain data into high- and low-creativity populations based on the mean human creativity rating assigned to each response. For each AUT stimulus, participant responses were rated on their creativity by four independent human raters ($ICC = 0.75$, good inter-rater agreement), and we followed a standard practice to aggregate these scores by averaging to yield a single creativity score per response (Organisciak et al., 2023; Amabile, 1982). Since responses are rated on a scale of 1 to 5 and the mean rating distribution is right-skewed (Appendix Figure 45), the choice of cutoff separating low- from high-creativity responses is not uniquely determined. We therefore report results across three cutoffs that roughly balance split sizes: 1.75, 2.0, and 2.25, where low- and high-creativity responses are those with $rating < cutoff$ and $rating \geq cutoff$ respectively. We use 2.0 in the main text and report the remaining two as robustness checks (Appendix C).

4 Experimental Setup

4.1 Models

We consider several open-source instruction following models varying in their size: GEMMA-3-{270M, 1B, 4B, 12B, 27B}-IT (Kamath et al., 2025), LLAMA-3.1-{8B, 70B}-INSTRUCT, LLAMA-3.2-{1B, 3B}-INSTRUCT (Grattafiori et al., 2024), OLMO-3-7B-INSTRUCT, OLMO-3.1-32B-INSTRUCT (Olmo et al., 2025), FALCON-40B-INSTRUCT (Almazrouei et al., 2023), DEEPSEEK-R1-DISTILL-LLAMA-70B (DeepSeek-AI, 2025), MISTRAL-7B-INSTRUCT-V0.3 (Jiang et al., 2023) and QWEN2.5-{0.5B, 1.5B, 3B, 14B, 32B, 72B}-INSTRUCT (Team, 2024; Yang et al., 2024a). Additionally, to study the alignment difference between high vs. low creativity responses and the effect of post-training techniques, we consider several variants of LLAMA-3.1-8B-INSTRUCT: LLAMA-3.1-8B which is the base pretrained model, CRPO-LLAMA-3.1-8B-INSTRUCT-CRE that has been trained to align with human creative preferences and optimized for multiple dimensions of creativity, namely, novelty, surprise, diversity, and quality (Ismayilzada et al., 2025), and DEEPSEEK-R1-DISTILL-LLAMA-8B that has been finetuned with the reasoning outputs of DEEPSEEK-R1 (DeepSeek-AI, 2025).

For the effects of the post-training objective analysis, to isolate the contribution of the specific creativity optimization technique from that of the training data and the fine-tuning paradigm, we additionally consider two ablations that share both the base model and the training data with CRPO-LLAMA-3.1-8B-INSTRUCT-CRE: CRPO-SFT-LLAMA-3.1-8B-INSTRUCT, which is LLAMA-3.1-8B-INSTRUCT supervised-finetuned on the MuCE creativity dataset, and CRPO-DPO-LLAMA-3.1-8B-INSTRUCT, which is the same base preference-tuned on MuCE using vanilla DPO (Ismayilzada et al., 2025).

Finally, to test whether the observed post-training effects generalize beyond a single base model family, we replicate the post-training analysis on a set of QWEN2.5 models: QWEN2.5-7B (base) and QWEN2.5-7B-INSTRUCT serve as matched within-family baselines, QWEN2.5-MATH-7B is finetuned on a mathematics corpus (Yang et al., 2024b), and DEEPSEEK-R1-DISTILL-QWEN-7B further finetunes QWEN2.5-MATH-7B with DEEPSEEK-R1 reasoning traces (DeepSeek-AI, 2025).

To ensure high creativity and diversity of outputs, we sample four generations for each prompt with $temperature=0.7$ and $top_p=0.95$.

4.2 Alignment Factors

To analyze the factors driving brain-LLM alignment, we compute the Pearson correlation between alignment scores and two model attributes: parameter count and performance on the AUT task. For evaluating AUT performance, we employ the GEMINI-3-FLASH (Google, 2025) model as LLMs have been shown to accurately predict human creativity on the AUT (Saretzki & Benedek, 2026; Organisciak et al., 2023). We slightly adapt the original prompt by instructing the model to produce only a few words, ensuring that it outputs a single, most original idea without additional text. This adjustment makes the outputs more comparable to human responses, which are typically concise, and enables consistent scoring (Appendix Tables H.1, H.2)

4.3 Model Representations

Since the model representations are extracted per token, past work has considered two main pooling strategies: extracting representations from the *last token* or averaging over all token representations (i.e., *mean token*). We consider both approaches in our work and indicate the type of pooling used for each analysis.

Because the AUT prompt instructs the model to *think of* creative uses, our models produce reasoning-like chains in their responses rather than a bare answer. Outputs from DEEPSEEK-R1-DISTILL variants additionally contain explicit `<think>...</think>` reasoning traces. At the response stage, we extract activations from the full set of response tokens, so the resulting representations capture this reasoning content rather than the final answer alone. Generation is capped at 1024 tokens for all models; per-model generation length statistics are reported in Appendix Table 4.

While we hold prompt content fixed in our primary analyses (Section 3.2), we additionally evaluate two baseline conditions that remove the task instruction: an *empty-prompt* condition, in which the AUT prompt is replaced by an empty input, and a *placeholder-prompt* condition, in which it is replaced by a sequence of # characters matched to its token length.

5 Results

LLMs Align with the Human Brain During Creative Thinking Figure 2 shows the brain alignment between the Default Mode Network (DMN) and model representations extracted during early cue processing — that is, after the prompt has been presented but before any response has been generated. Notably, the DMN has been directly implicated in creative cognition and is considered a core neural substrate of creative thinking (Luchini et al., 2025). We observe moderate and consistent alignment across models, and crucially, this **alignment is positively correlated with both model size** ($r = 0.55, p < 0.05$) and **creative task performance** as measured by the AUT score ($r = 0.52, p < 0.05$). This suggests that larger and more creatively capable models more closely mirror the neural representations underlying human divergent thinking during the initial stages of ideation. We note that the two relationships emerge under different pooling strategies: the size correlation is obtained with last-token pooling and the AUT-score correlation with mean-token pooling, with each attribute showing a weaker relationship under the alternative pooling (Appendix Figures 12, 13).

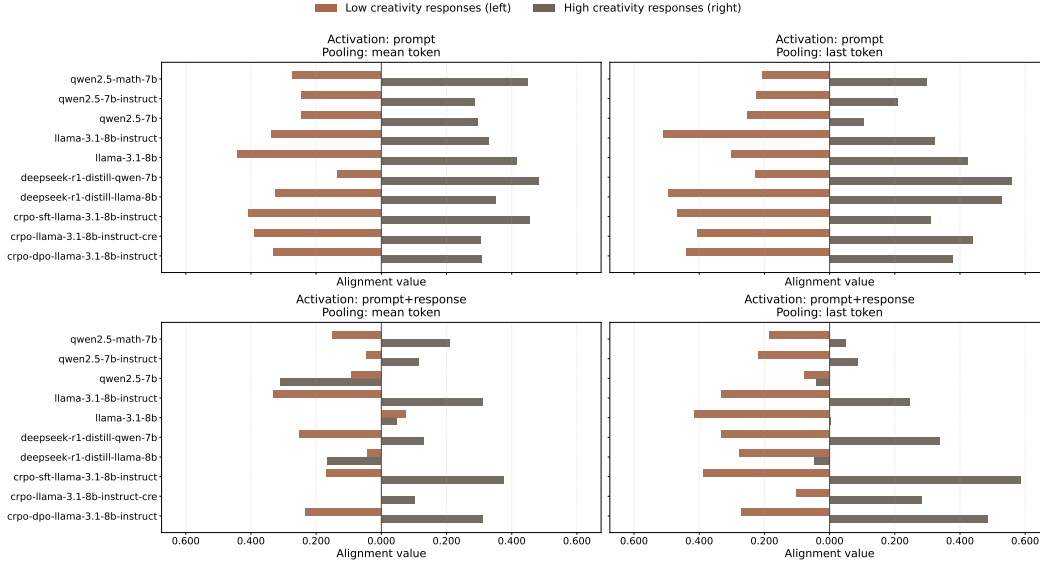


Figure 4: Default Mode Network (DMN) AUT brain alignment results for high and low creativity response populations (rating cutoff = 2.0). The left red and right gray bars indicate positive alignment with high and low creativity responses, respectively. Flipped bars indicate negative alignment. Results for rating cutoffs 1.75 and 2.25 can be found in Appendix Figures 10 and 11 respectively.

However, when alignment is measured using representations extracted after the model has generated its response, alignment values become more variable across models, and the correlations with both model size and task performance weaken (Appendix Figures 14, 15). This decoupling may reflect the fact that models tend to converge on similar responses regardless of scale (Wenger & Kenett, 2025; Ismayilzada et al., 2024b), or alternatively, that generated responses increasingly diverge from human responses in length, structure, and quality as model scale grows. We note that this weakening is observed when pooling across model families: within the Gemma-3 family, where pretraining data and architecture are held approximately fixed, the size and AUT-score correlations remain positive at the response stage (Appendix E). This suggests that the loss of correlation at the response stage may partly reflect heterogeneity across families rather than a property of the generation stage itself. We also replicate some of these patterns (i.e. correlation with AUT score) for the Frontoparietal Network (FPN) (Appendix Figure 17). To assess whether these effects are specific to creative cognition or reflect more general properties of model or brain representations, we repeat the alignment analyses using the non-creative OCT task and a brain network not typically associated with creativity, namely the Somatomotor network (SOM) (Ji et al., 2019). We find that neither model size nor task performance correlates significantly with alignment in either of these conditions across pooling strategies (Appendix Figures 20-25), providing an important double dissociation that strengthens the interpretation that the observed alignment effects are driven specifically by the conjunction of a creativity-relevant task and a creativity-relevant brain network. We further verify that these effects are not attributable to differences in generation length across models. Although generated response length varies across our model set (Appendix Table 4), neither mean nor median generation length correlates significantly with DMN alignment during the AUT (Appendix Figures 39-42), indicating that the alignment patterns we report do not simply track how much text each model produces. As a complementary analysis, we also compare absolute alignment levels across these conditions (Appendix G), which are reliably higher in the DMN than in the SOM at the prompt stage, but comparable between the AUT and OCT conditions in most configurations.

Finally, we examine the two baseline conditions that remove the task instruction (Section 3.2). Neither the *empty-prompt* nor the *placeholder-prompt* condition yields significant correlations

between alignment and model size or task performance, across any combination of stage and token pooling (Appendix Figures 26-33). Alignment in the creativity-relevant networks therefore depends on the model receiving the same task instructions given to human participants, rather than arising from generic input-encoding effects.

Peak Alignment Depth Depends on Token Pooling Different layers of a language model capture qualitatively different aspects of linguistic and cognitive processing: earlier layers tend to encode lexical and syntactic properties, while later layers are associated with more abstract, task-specific representations (Tenney et al., 2019; Jawahar et al., 2019), and prior work has linked these layer-wise differences to brain alignment directly (Oota et al., 2023; He et al., 2026). We therefore asked at what depth of the network alignment with creativity-relevant brain networks peaks, computing for each model the relative depth of its best-aligned layer (best layer index divided by total number of layers).

Figure 3 shows that the answer depends on how token representations are pooled. Under last-token pooling, peak-alignment depths are modal in the earliest portion of the network (relative depth ≈ 0.15 – 0.25) and otherwise spread broadly across the remaining depths. Under mean-token pooling, the distribution is bimodal, with a pronounced mode in the final layers (relative depth > 0.85) and a smaller secondary mode at the earliest depths, and comparatively little mass in between. We further find no reliable relationship between the relative depth at which alignment peaks and the magnitude of that alignment (Appendix Figures 35-38). Taken together, these results indicate that the depth at which LLM representations best match creativity-related neural geometry is not a stable property of the model, but is contingent on the representation-extraction choice.

Finally, we note that our layer-selection procedure takes the maximum alignment across all layers of each model (Schrimpf et al., 2020), which yields upper bounds rather than unbiased estimates; we quantify this optimism bias via cross-validated layer selection in Appendix H, and find that it does not scale with layer count and that the size correlation reported above survives its removal.

Post-Training Objectives Are Associated with Differential Alignment to High- and Low-Creativity Neural Responses To examine how different post-training approaches relate to brain alignment, we partition the brain data into high and low creativity populations as described in Section 3.4 and measure alignment separately for each population across variants of LLAMA-3.1-8B and QWEN2.5-7B (Section 4.1). Results are shown in Figure 4, which reports the 2.0 rating cutoff; the same patterns largely hold under the 1.75 and 2.25 cutoffs (Appendix Figures 10 and 11). During early prompt processing, all model variants show broadly comparable alignment with both populations. Once models begin generating responses, however, the alignment patterns diverge across training conditions. Comparing LLAMA-3.1-8B-INSTRUCT with its creativity-optimized counterpart CRPO-LLAMA-3.1-8B-INSTRUCT-CRE, the instruct model shows positive alignment with both populations, whereas the CrPO model shows positive alignment with the high-creativity population but no reliable positive alignment with the low-creativity one, a selective pattern consistent with its training objective. DEEPSEEK-R1-DISTILL-LLAMA-8B exhibits the opposite asymmetry to CrPO: negative alignment with high-creativity responses and no reliable negative alignment with low-creativity ones, consistent with reasoning-chain training being associated with representations further from the neural geometry of creative ideation.

Two additional comparisons help localize these effects. First, the CrPO pattern appears specific to the full optimization method rather than to its training data or fine-tuning paradigm: CRPO-SFT-LLAMA-3.1-8B-INSTRUCT and CRPO-DPO-LLAMA-3.1-8B-INSTRUCT, which share both the base model and the MuCE dataset with CRPO-LLAMA-3.1-8B-INSTRUCT-CRE, both show positive alignment with the low-creativity population, and only the full CrPO variant lacks it. Second, the reasoning-training pattern partially generalizes beyond the Llama family: DEEPSEEK-R1-DISTILL-QWEN-7B and QWEN2.5-MATH-7B both show positive alignment with the low-creativity population, while the vanilla QWEN2.5-7B and QWEN2.5-7B-INSTRUCT baselines show no consistent lean toward either population.

6 Discussion

In summary, our results show that 1) LLM alignment with the human brain is positively associated with model size and with performance on divergent thinking assessments, 2) the depth at which alignment peaks depends on the token-pooling strategy, and does not predict alignment magnitude, and 3) post-training objectives are associated with systematic differences in alignment, with the largest elevations observed for objectives explicitly designed to push models toward greater creativity in responses. Notably, these patterns are stage-dependent: the relationship between model properties and brain alignment is clearest during prompt processing and weakens once models begin generating responses, suggesting that the representational dynamics of language generation diverge from human creative cognition in ways that are not simply resolved by scaling. Our post-training results further indicate that the choice of training objective is associated with selective differences in how model representations relate to the neural geometry of creative thought: reasoning-chain training is associated with negative alignment to high-creativity neural responses, while creativity-optimized training shows the opposite asymmetry, retaining alignment with high-creativity responses while lacking the positive low-creativity alignment.

Our post-training results carry implications for how language models are fine-tuned in future work. Post-training strategies and datasets have thus far emphasized tasks requiring *convergent thinking*, such as solving difficult math problems or coding tasks (DeepSeek-AI, 2025; Lightman et al., 2023; Chen et al., 2021), because they are easily evaluated using a human-provided ground truth or an LLM judge. While this has proven effective for developing models that can solve tasks which emphasize convergent thinking — reasoning toward a single correct answer or solution — our findings raise the possibility that it may simultaneously come at a cost to the ability of LLMs to come up with many and varied ideas, the hallmark of divergent thinking (Runco & Acar, 2012). Specifically, the reasoning-trained models in our set showed alignment patterns shifted toward low-creativity neural responses relative to their creativity-optimized counterparts, consistent with the possibility that reasoning-oriented post-training moves representations toward the neural geometry associated with lower creativity.

Together, these findings suggest that brain alignment may offer an informative lens for evaluating and developing language models in the context of creative cognition — one that goes beyond behavioral benchmarks and captures something closer to the computational principles underlying human creative thought. Given the crucial role of creativity in fostering scientific, engineering, and artistic advances (Torrance, 1974; Guilford, 1967), developing LLMs capable of creativity when tasked with challenging, long-horizon, or ambiguously scoped problems is an important goal. Our findings suggest that post-training exclusively with convergent-thinking tasks warrants closer scrutiny for its effects on divergent thinking.

Conclusion

We presented the first investigation of brain-LLM alignment in the context of divergent creative thinking, using fMRI data from participants performing the Alternative Uses Test alongside a non-creative control task. Our results reveal that alignment with creativity-relevant brain networks is positively associated with model size and creative task performance during prompt processing, but that this relationship weakens once models generate responses, suggesting that the generative process introduces representational dynamics that diverge from human creative cognition. We further find that post-training objectives are associated with selective differences in alignment: creativity-optimized training retains alignment with high-creativity neural responses while lacking the positive low-creativity alignment present in other variants, whereas reasoning-chain training shows the opposite asymmetry, with negative alignment to high-creativity responses. Together, these findings suggest that the correspondence between LLM representations and the neural geometry of human creative thought is nuanced and sensitive to both the stage of processing examined and the nature of the training objective. We hope these results encourage further investigation into the neural plausibility of creative cognition in language models.

Acknowledgements

LP gratefully acknowledges the support of the Swiss National Science Foundation (grant 205121_207437: C - LING). R.E.B. is supported by Amazon and grants from the National Science Foundation (DRL-2400782; DUE-2155070). AB also gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), the European Research Council (Starting grant no. 101222478, RESPECT-LM), the AI2050 program at Schmidt Sciences (Grant #G-25-69783), Sony Group Corporation, and the Swiss National Supercomputing Center (CSCS) in the form of an infrastructure engineering and development project.

References

- Anna Abraham. The promises and perils of the neuroscience of creativity. *Frontiers in human neuroscience*, 7:246, 2013.
- Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24332–24350, 2025.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.
- Teresa M Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997, 1982.
- Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*, 2023.
- Roger E Beaty, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative cognition and brain network dynamics. *Trends in cognitive sciences*, 20(2):87–95, 2016.
- Roger E Beaty, Yoed N Kenett, Alexander P Christensen, Monica D Rosenberg, Mathias Benedek, Qunlin Chen, Andreas Fink, Jiang Qiu, Thomas R Kwapil, Michael J Kane, et al. Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, 115(5):1087–1092, 2018.
- Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026.
- Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.
- Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.

- Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.
- K. Duncker and L.S. Lees. *On Problem-solving*. Psychological monographs. American Psychological Ass., 1948. URL <https://books.google.ch/books?id=g888tAEACAAJ>.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.
- Luis Fabricio Góes, Marco Volpe, Piotr Sawicki, Marek Grses, and Jacob Watson. Pushing gpt’s creativity to its limits: Alternative uses and torrance tests. 2023.
- Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *arXiv preprint arXiv:2411.05712*, 2024.
- 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.
- Team Google. A new era of intelligence with gemini 3, 2025. URL <https://blog.google/products-and-platforms/products/gemini/gemini-3/#note-from-ceo>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Joy Paul Guilford. The nature of human intelligence. 1967.
- Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. *Advances in Neural Information Processing Systems*, 38:70967–70989, 2026.
- Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*, 2024a.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*, 2024b.
- Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. Creative preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9580–9609, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.509. URL <https://aclanthology.org/2025.findings-emnlp.509/>.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL <https://aclanthology.org/P19-1356/>.
- Jie Lisa Ji, Marjolein Spronk, Kaustubh Kulkarni, Grega Repovš, Alan Anticevic, and Michael W Cole. Mapping the human brain’s cortical-subcortical functional network organization. *NeuroImage*, 185:35–57, 2019.
- Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 10(3), 2023.

- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4, 2025.
- Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.
- Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.
- Takuya Koumura, Hiroki Terashima, and Shigeto Furukawa. Human-like modulation sensitivity emerging through optimization to natural sound recognition. *Journal of Neuroscience*, 43(21):3876–3894, 2023.
- Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.
- Elisa Kwon, John D Patterson, Roger E Beaty, and Kosa Goucher-Lambert. Assessing the alignment between word representations in the brain and large language models. In *International Conference on Design Computing and Cognition*, pp. 207–223. Springer, 2024.
- Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The twelfth international conference on learning representations*, 2023.
- Cheng Liu, Kaixiang Zhuang, Daniel C Zeitlen, Qunlin Chen, Xueyang Wang, Qiuyang Feng, Roger E Beaty, and Jiang Qiu. Neural, genetic, and cognitive signatures of creativity. *Communications Biology*, 7(1):1324, 2024.
- Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2776–2794, 2025.
- Simone A Luchini, Emmanuelle Volle, and Roger E Beaty. The role of the default mode network in creativity. *Current Opinion in Behavioral Sciences*, 65:101551, 2025.
- Paul M Matthews and Peter Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004.
- Sarnoff Mednick. The associative basis of the creative process. *Psychological review*, 69(3):220, 1962.
- Manh Hung Nguyen and Adish Singla. Divergent-convergent thinking in large language models for creative problem generation. *arXiv preprint arXiv:2512.23601*, 2025.
- Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heine-man, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah,

- Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL <https://arxiv.org/abs/2512.13961>.
- Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneeesh Singh, Raju Surampudi Bapi, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). In *International Conference on Learning Representations*, volume 2025, pp. 91759–91788, 2025.
- SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36:18001–18014, 2023.
- Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356, 2023.
- Christopher Pinier, Sonia Acuña Vargas, Mariia Steeghs-Turchina, Dora Matzke, Claire E Stevenson, and Michael D Nunez. Large language models show signs of alignment with human neurocognition during abstract reasoning. *arXiv preprint arXiv:2508.10057*, 2025.
- Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. Fine-tuning on diverse reasoning chains drives within-inference cot refinement in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3789–3808, 2025.
- Mark A Runco and Selcuk Acar. Divergent thinking as an indicator of creative potential. *Creativity research journal*, 24(1):66–75, 2012.
- Janika Saretzki and Mathias Benedek. Investigating the validity evidence of automated scoring methods for divergent thinking assessments. *Creativity Research Journal*, pp. 1–17, 2026.
- Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2020.
- Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.
- Shashank Srikant, Ben Lipkin, Anna Ivanova, Evelina Fedorenko, and Una-May O’Reilly. Convergent representations of computer programs in human and artificial neural networks. *Advances in Neural Information Processing Systems*, 35:18834–18849, 2022.
- Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3’s creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*, 2022.
- Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

- Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5303–5324, 2024.
- E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1974.
- Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.
- Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.
- Emily Wenger and Yoed Kenett. We’re different, we’re the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*, 2025.
- Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.
- BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

A Limitations

Correlational rather than causal evidence. We emphasize that our findings are correlational rather than causal. The model variants we compare differ along multiple dimensions beyond their post-training objective — including pretraining data, optimizer, fine-tuning corpus, and generation length — and a causal attribution would require training a single fixed base model under each objective with matched data and generation budgets, which is beyond the scope of the present work. Several of the patterns we report also hold only partially across model families: the negative high-creativity alignment observed for the Llama reasoning distill is not reproduced in its Qwen counterpart, where only the positive low-creativity alignment replicates. Our claims should therefore be read as identifying systematic associations between post-training objectives and brain alignment, not as establishing that these objectives cause the observed representational shifts.

No mechanistic account of the observed effects. Our analyses establish that different post-training objectives are associated with different alignment patterns, but they do not explain why. One behavioral explanation could be that creativity-optimized models produce responses rated more creative than those of reasoning-distilled models. However, a full mechanistic account would require identifying where and how representations diverge from their shared base model: layer-wise comparisons localizing the divergence within the network, probing analyses testing whether creativity-relevant features are linearly decodable from model activations, or component-level analyses attributing the shift to specific attention heads or MLP blocks. We regard these as the natural next steps.

Instruction tuning and reasoning budget are not isolated. Two design factors remain uncontrolled in our setup. First, we restrict our analyses to instruction-tuned models, as the AUT is an active task requiring instruction following, and rely on prior work (Aw et al., 2023) for the effect of instruction tuning itself; we do not report a within-model comparison of the same base model before and after instruction tuning, nor a layer-wise account of how tuning transforms generic prompt representations into task-specific ones. Second, our generation setup caps output at 1024 tokens but does not control the reasoning budget: the models in our set are non-thinking models by default, with the exception of the reasoning-distilled variants, and we do not include models with an explicit thinking mode. A systematic comparison of thinking versus non-thinking models under matched reasoning budgets would be required to disentangle these factors.

Finally, our post-training comparison does not include a model fine-tuned for domain-general cognitive modelling, which would provide a useful reference point against the creativity- and reasoning-optimised objectives we do examine.

B RSA Predictivity

To compute the RSA predictivity, we compare the geometry of representational dissimilarity matrices (RDMs) constructed from LLM representations and fMRI responses, respectively. LLM RDMs are constructed using Pearson correlation distance ($1 - r$) between pairs of stimulus representations. On the brain side, we use all voxels within each Yeo parcel without further voxel selection; each subject’s brain is resampled into the same MNI space as the Yeo parcellation, so no additional processing is required to apply the parcellation to individual brains. RDMs are constructed per subject rather than pooled across subjects, and the two RDMs are compared via Spearman rank correlation between their upper triangles. Because this comparison metric is signed, alignment scores range over $[-1, 1]$, with negative values indicating anti-alignment with the target population’s representational geometry. Following Nili et al. (2014), each subject’s noise ceiling is the Spearman correlation between that subject’s RDM and the mean RDM computed from the remaining $N - 1$ subjects.

C Detailed Post-training Results

Tables 1, 2 and 3 provide raw alignment medians with 95% bootstrapped CIs for both token pooling modes (last-token, mean-token) and all three rating cutoff thresholds: 1.75 ($N_{low} = 1,594$, $N_{high} = 2,034$), 2.0 ($N_{low} = 2,147$, $N_{high} = 1,481$), and 2.25 ($N_{low} = 2,588$, $N_{high} = 1,040$). The “alignment” column marks each cell as + (CI lies largely on the positive side), − (CI lies largely on the negative side), or $\pm$ otherwise; we use ± 0.1 as the practical boundary since alignment values within 0.1 of zero are negligible. Prompt-stage results are omitted for brevity (all $\pm$).

D Robustness of the Alignment Correlations

The brain alignment correlations are computed over $n = 24$ models. At this sample size, Pearson correlations can be influenced by individual observations, and a single point estimate with an uncorrected p -value understates this fragility. We therefore compute three additional diagnostics for every cell in our analysis grid: a leave-one-out (LOO) jackknife over models, the Spearman rank correlation alongside Pearson r , and a bias-corrected and accelerated (BCa) bootstrapped 95% confidence interval. Below we summarize our findings:

Primary correlations. The DMN–size correlation at the prompt stage is robust across all three diagnostics: $r = +0.55$, BCa CI $[+0.24, +0.77]$, LOO range $[+0.45, +0.61]$, and Spearman $\rho = +0.65$ ($p < 0.001$). The rank structure is therefore at least as strong as the linear fit, and no single model carries the effect: removing the most influential observation (LLAMA-3.1-70B-INSTRUCT) reduces r only to $+0.45$. The AUT-score correlations are positive but less well estimated. In the DMN, $r = +0.52$ with BCa CI $[+0.07, +0.76]$ and LOO range $[+0.40, +0.58]$; in the FPN, $r = +0.46$ with a BCa CI that straddles zero, $[-0.06, +0.77]$, and LOO range $[+0.29, +0.59]$. In both cases Spearman ρ is substantially lower than Pearson r ($+0.26$ and $+0.22$, neither reaching significance), indicating that the linear fit is carried in part by magnitude differences that the rank correlation does not capture.

Leverage point in the AUT-score correlations. The LOO analysis identifies GEMMA-3-270M-IT as the influential observation for both AUT-score correlations. It is the smallest model in our set, so it anchors the low end of both the size and AUT-score ranges. Removing it shifts the DMN correlation from $r = +0.52$ to $+0.40$ and the FPN correlation from $+0.46$ to $+0.29$, preserving the positive direction while reducing the magnitude. A denser sampling of small models would likely stabilise these estimates.

Multiple comparisons. Our full grid spans model $\times$ network $\times$ brain task $\times$ model task $\times$ stage $\times$ token pooling, and reporting uncorrected p -values across all cells would inflate the false-positive rate. We pre-specify a primary family of three correlations corresponding to the positive effects claimed in the main text: alignment against model size and AUT score at the prompt stage in the matched AUT condition, in the two creativity-relevant networks. After Bonferroni correction over this family, the DMN–size ($p_{\text{bonf}} = 0.016$) and DMN–AUT-score ($p_{\text{bonf}} = 0.036$) correlations remain significant, while the FPN–AUT-score correlation does not ($p_{\text{bonf}} = 0.079$).

Control conditions. The control cells behave as predicted. In the matched non-creative OCT condition, the DMN–size correlation at the prompt stage is $r = +0.09$ ($p = 0.66$) with last-token pooling and $r = +0.22$ ($p = 0.30$) with mean-token pooling, compared with $r = +0.55$ ($p = 0.005$) in the matched AUT condition; no FPN–OCT cell reaches significance ($-0.34 \leq r \leq +0.23$, all $p > 0.10$). The FPN–size correlations in the AUT condition, which we do not claim as a positive effect, are likewise null across all four stage and pooling combinations ($-0.27 \leq r \leq +0.05$, all $p > 0.20$). Of the eight Somatomotor control cells, seven are non-significant; the exception is a negative correlation with model size at the response stage under mean-token pooling ($r = -0.42$, $p = 0.046$), which is opposite in direction to the effects we report.

E Within-Family Scaling

The correlations reported in the main text pool models across families, which confounds parameter count with differences in pretraining data, tokenizer, and architecture. As a complementary analysis that holds these factors approximately fixed, we report within-family scaling scans for the two families in our set with sufficient size coverage: Gemma-3 (270M–27B, $n = 5$) and Qwen2.5 (0.5B–72B, $n = 6$). All correlations are computed in the DMN on the AUT task, using last-token pooling for model size and mean-token pooling for AUT score, matching the main analyses.

Gemma-3. Alignment increases monotonically with model size across the Gemma-3 family at both stages (Figures 5 and 6). At the prompt stage, $r = 0.90$ ($p = 0.04$) for model size and $r = 0.85$ ($p = 0.07$) for AUT score; at the prompt+response stage, $r = 0.95$ ($p = 0.01$) and $r = 0.97$ ($p = 0.01$) respectively.

Qwen2.5. The Qwen2.5 family shows the same positive scaling relationship at the prompt stage (Figure 7): $r = 0.89$ ($p = 0.02$) for model size and $r = 0.74$ ($p = 0.09$) for AUT score. We note that parameter counts in this family are unevenly distributed, with four of six models below 10B and the remaining two at 32B and 72B, so the size correlation is sensitive to the two largest models.

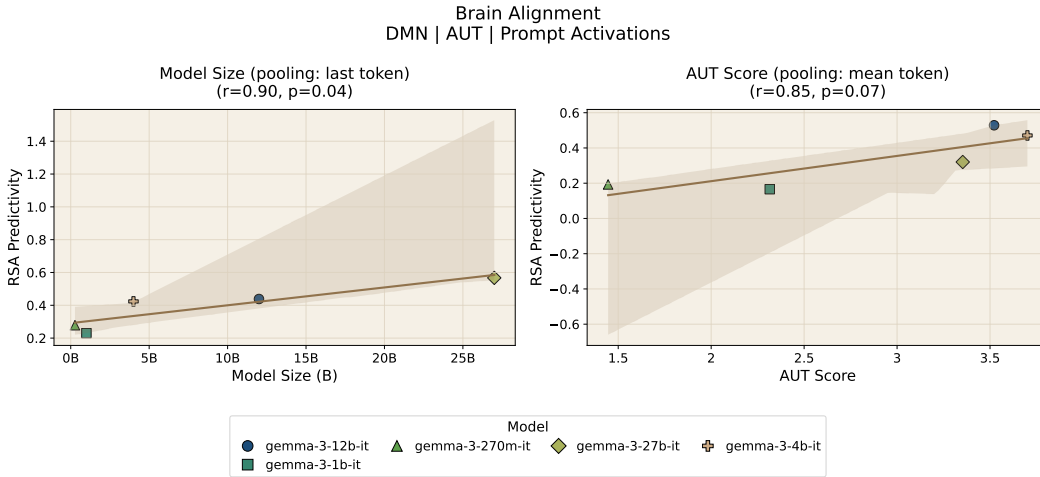


Figure 5: Default Mode Network (DMN) AUT Gemma family brain alignment results by model size and task performance using prompt-only model activations. r and p correspond to the Pearson correlation coefficient and p-value.

F Per-Stimulus and Per-Subject Sample Sizes for the Response Creativity Partitioning

Because alignment is computed per subject and aggregated by taking the median across subjects (Section 3.3), the high/low creativity partition can in principle produce subjects or stimuli that are represented in only one of the two populations. We therefore report the distribution of responses across the two buckets at both levels. Figures 8 and 9 show these distributions at the 2.0 cutoff used in the main text.

Per-stimulus distribution. Figure 8 shows the number of low- and high-creativity responses for each of the 46 AUT stimuli, sorted by the proportion of high-creativity responses. Total responses per stimulus range from approximately 61 to 94, and every stimulus contributes responses to both buckets. The proportion of high-creativity responses varies from roughly 0.66 (*comb*, *candle*, *balloon*) to roughly 0.22 (*bucket*, *dog leash*, *brick*), indicating that

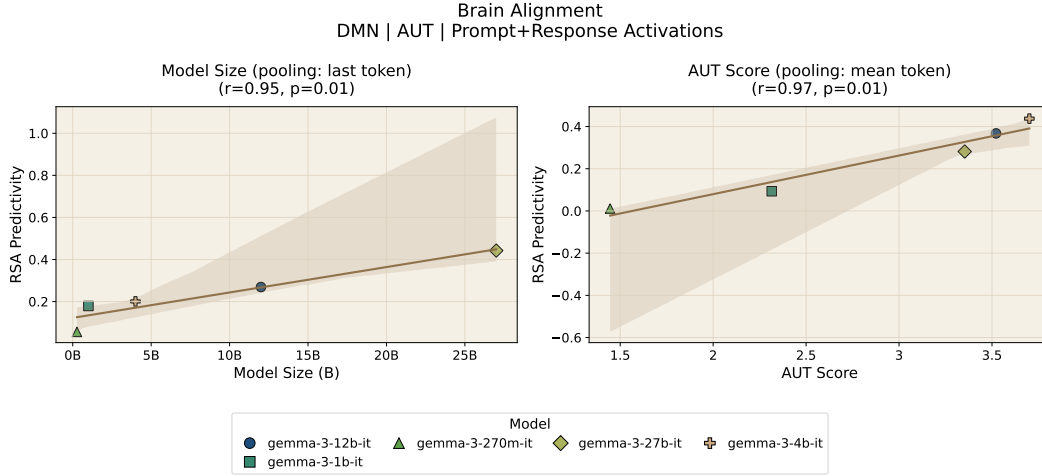


Figure 6: Default Mode Network (DMN) AUT Gemma family brain alignment results by model size and task performance using prompt+response model activations. r and p correspond to the Pearson correlation coefficient and p-value.

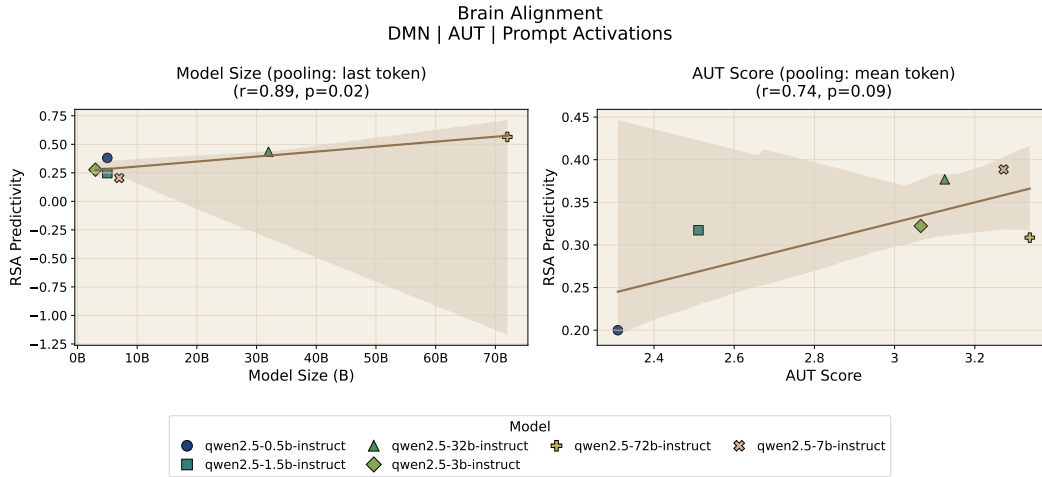


Figure 7: Default Mode Network (DMN) AUT Qwen family brain alignment results by model size and task performance using prompt-only model activations. r and p correspond to the Pearson correlation coefficient and p-value.

some objects more readily elicit original uses than others, but that no stimulus is confined to a single bucket. The partition is therefore not driven by a subset of stimuli falling entirely on one side of the cutoff.

Per-subject distribution. Figure 9 plots, for each subject, the number of low-creativity responses against the number of high-creativity responses. Because each subject completed a fixed number of AUT trials, the points lie along an anti-diagonal band rather than scattering freely: a subject with more low creativity responses necessarily has fewer high-creativity ones. Most subjects fall below the identity line, consistent with the overall imbalance at this cutoff ($N_{low} = 2,147$ vs. $N_{high} = 1,481$).

The relevant caveat is visible at the extremes of this distribution. While the majority of subjects contribute between 5 and 18 responses to each bucket, a small number contribute very few to one of them: several subjects have fewer than three high-creativity responses, and at least one has none. For these subjects, the RDM in the sparser bucket is estimated

from very few stimuli and the corresponding alignment value is correspondingly noisy. Because we aggregate using the median rather than the mean, these subjects have limited influence on the reported values, but we note their presence explicitly.

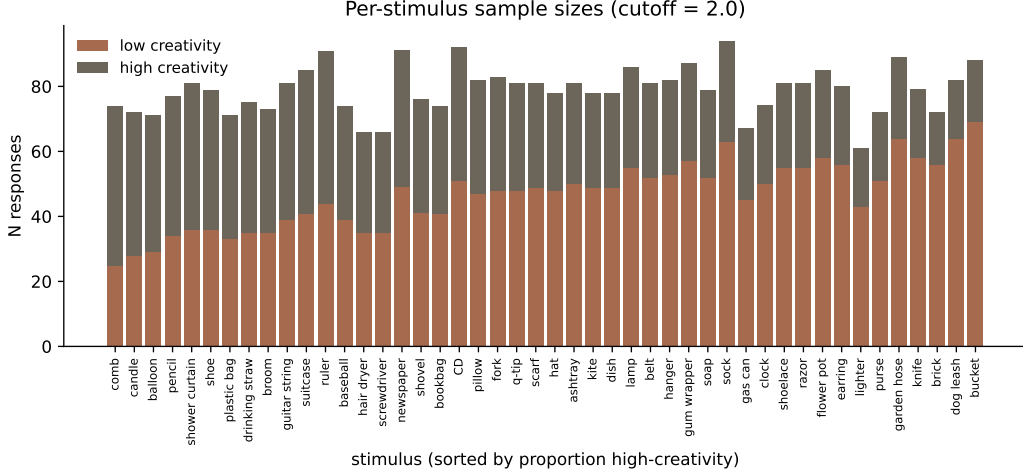


Figure 8: Number of low- and high-creativity responses per AUT stimulus at the 2.0 cutoff, sorted by the proportion of high-creativity responses. All 46 stimuli contribute to both buckets.

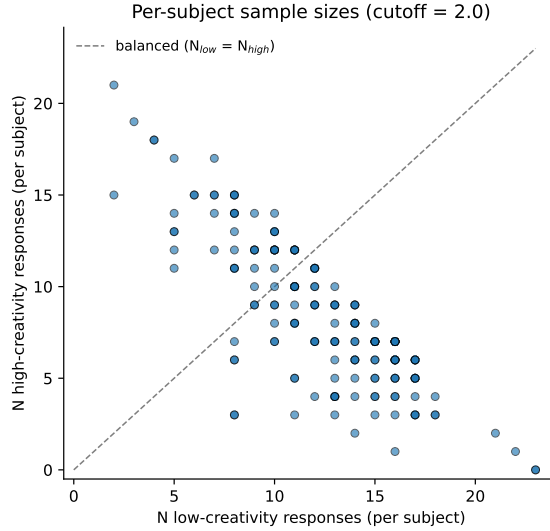


Figure 9: Per-subject sample sizes in the low- and high-creativity buckets at the 2.0 cutoff. Each point is one subject; the dashed line marks equal counts. The anti-diagonal band reflects the fixed number of AUT trials per subject.

G Absolute Alignment Levels Across Conditions

Our main analyses concern how alignment relates to model properties. As a complementary descriptive analysis, we also compare absolute alignment levels across conditions. Since the same models are evaluated in every condition, all comparisons are paired over models: we compute one difference per model and test it against zero with a Wilcoxon signed-rank test, with a sign test as a distribution-free cross-check and a bootstrap CI on the median

difference. Twelve comparisons are reported; we note where results survive Bonferroni correction across them.

Creativity-relevant versus control network. Comparing the DMN with the Somatomotor control network on the AUT task, alignment is reliably higher in the DMN at the prompt stage under both pooling strategies: median difference +0.15 (CI [+0.12, +0.21]; 18 of 23 models; $p = 0.002$) with last-token pooling, and +0.16 (CI [+0.07, +0.27]; 22 of 23 models; $p < 10^{-6}$) with mean-token pooling. Both survive correction. Neither response-stage cell shows a reliable difference ($p = 0.94$ and $p = 0.34$).

Creative versus non-creative task. The corresponding comparison between the matched AUT and OCT conditions is less consistent. Of eight cells, one shows a clear AUT advantage and survives correction (DMN, prompt stage, mean-token pooling: median difference +0.14, CI [+0.09, +0.18]; 23 of 24 models; $p < 10^{-6}$). Two are nominally significant in opposite directions and neither survives correction (FPN, response stage, last-token pooling: +0.12, $p = 0.034$; DMN, response stage, mean-token pooling: -0.11 , $p = 0.039$). The remaining five are null, including the DMN at the prompt stage under last-token pooling, where median alignment is 0.44 for AUT and 0.34 for OCT but models split evenly (12 of 24 in each direction, $p = 0.46$).

Summary. Absolute alignment in the DMN is reliably higher than in the Somatomotor control network at the prompt stage, while absolute alignment during the non-creative OCT task is comparable in magnitude to the creative AUT task in most configurations. These comparisons concern alignment magnitudes and are complementary to the correlational analyses reported in the main text, which concern the relationship between alignment and model properties.

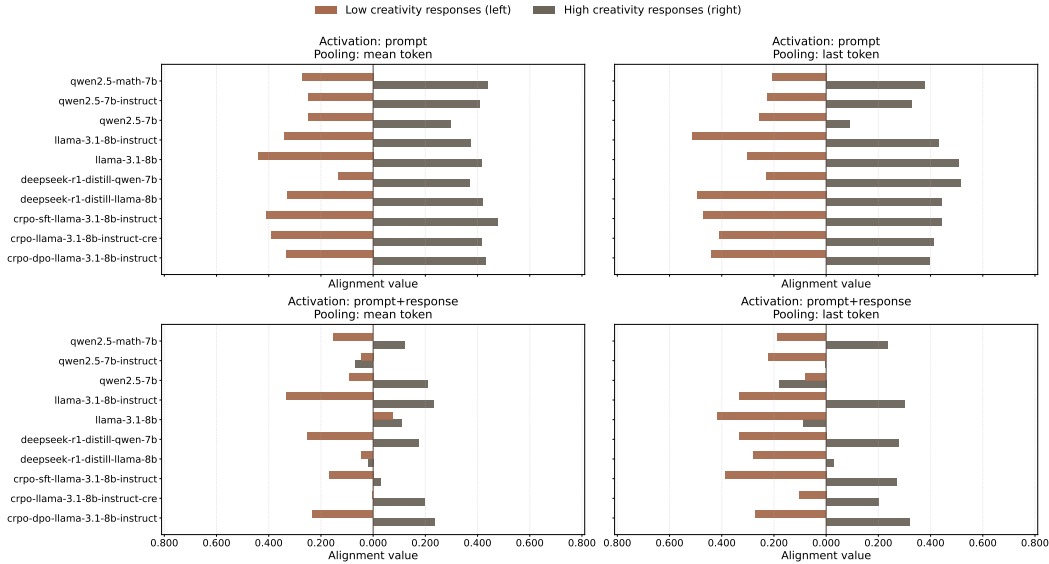


Figure 10: **Default Mode Network (DMN) AUT brain alignment results for high and low creativity response populations (rating cutoff = 1.75).** The left red and right gray bars indicate positive alignment with high and low creativity responses, respectively. Flipped bars indicate negative alignment.

H Cross-Validated Layer Selection

Our reported alignment scores take the maximum across layers of each model (Schrimpf et al., 2020). Because the expected maximum over more candidates is higher, this estimator

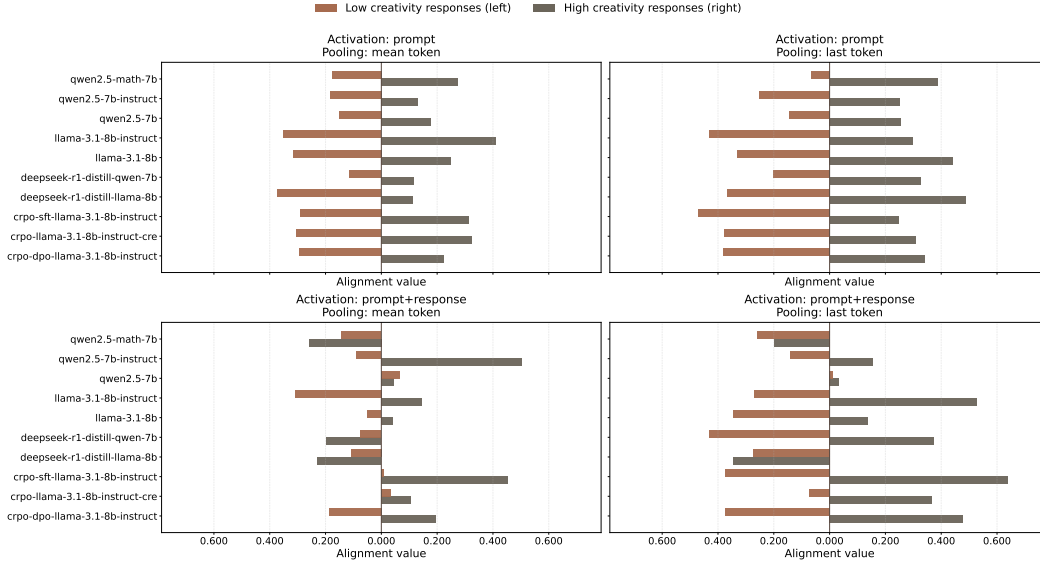


Figure 11: **Default Mode Network (DMN) AUT brain alignment results for high and low creativity response populations (rating cutoff = 2.25).** The left red and right gray bars indicate positive alignment with high and low creativity responses, respectively. Flipped bars indicate negative alignment.

model	pop	last-token			mean-token		
		median	95% CI	alignment	median	95% CI	alignment
crpo-dpo-llama-3.1-8b-instruct	high	+0.320	[-0.119,+0.507]	±	+0.104	[-0.208,+0.384]	±
crpo-dpo-llama-3.1-8b-instruct	low	+0.281	[+0.040,+0.530]	+	+0.159	[-0.127,+0.453]	±
crpo-llama-3.1-8b-instruct-cre	high	+0.158	[-0.200,+0.460]	±	+0.132	[-0.189,+0.415]	±
crpo-llama-3.1-8b-instruct-cre	low	+0.104	[-0.162,+0.336]	±	+0.007	[-0.206,+0.097]	—
crpo-sft-llama-3.1-8b-instruct	high	+0.270	[+0.010,+0.542]	+	+0.018	[-0.273,+0.200]	±
crpo-sft-llama-3.1-8b-instruct	low	+0.300	[+0.045,+0.623]	+	+0.124	[-0.294,+0.617]	±
deepseek-r1-distill-llama-8b	high	+0.027	[-0.173,+0.170]	±	-0.019	[-0.358,+0.117]	±
deepseek-r1-distill-llama-8b	low	+0.175	[-0.248,+0.404]	±	+0.027	[-0.328,+0.269]	±
deepseek-r1-distill-qwen-7b	high	+0.217	[+0.047,+0.556]	+	+0.173	[-0.109,+0.386]	±
deepseek-r1-distill-qwen-7b	low	+0.272	[+0.023,+0.555]	+	+0.259	[+0.075,+0.439]	+
llama-3.1-8b	high	-0.104	[-0.409,+0.276]	±	+0.073	[-0.297,+0.280]	±
llama-3.1-8b	low	+0.322	[-0.113,+0.596]	±	-0.153	[-0.394,+0.141]	±
llama-3.1-8b-instruct	high	+0.176	[-0.225,+0.404]	±	+0.233	[+0.003,+0.422]	+
llama-3.1-8b-instruct	low	+0.326	[+0.151,+0.461]	+	+0.291	[+0.054,+0.546]	+
qwen2.5-7b	high	-0.181	[-0.549,+0.193]	±	+0.210	[-0.128,+0.434]	±
qwen2.5-7b	low	+0.079	[-0.150,+0.240]	±	+0.081	[-0.190,+0.376]	±
qwen2.5-7b-instruct	high	-0.002	[-0.183,+0.307]	±	-0.066	[-0.298,+0.427]	±
qwen2.5-7b-instruct	low	+0.182	[-0.055,+0.343]	+	+0.050	[-0.213,+0.237]	±
qwen2.5-math-7b	high	+0.235	[-0.020,+0.444]	+	+0.119	[-0.160,+0.476]	±
qwen2.5-math-7b	low	+0.192	[+0.001,+0.523]	+	+0.114	[-0.082,+0.306]	+

Table 1: Default Mode Network (DMN) AUT alignment results (prompt+response activations) per model and creativity population with rating cutoff=1.75.

is biased upward, and in principle the bias could scale with layer count. Since layer count is strongly collinear with parameter count in our set ($r = 0.94$), such a bias could in principle produce a size–alignment correlation with no representational basis. We therefore quantify it directly.

Procedure. For each model we hold out subjects when selecting the layer. Subjects are partitioned into five folds; for each fold, the best layer is chosen as the argmax of the median alignment across the remaining four folds, and alignment is then evaluated on the held-out

model	pop	last-token			mean-token		
		median	95% CI	alignment	median	95% CI	alignment
crpo-dpo-llama-3.1-8b-instruct	high	+0.307	[+0.023,+0.673]	+	+0.312	[+0.114,+0.620]	+
crpo-dpo-llama-3.1-8b-instruct	low	+0.348	[+0.040,+0.726]	+	+0.159	[-0.167,+0.453]	±
crpo-llama-3.1-8b-instruct-cre	high	+0.284	[-0.050,+0.495]	+	+0.080	[-0.125,+0.244]	±
crpo-llama-3.1-8b-instruct-cre	low	+0.077	[-0.256,+0.298]	±	+0.007	[-0.271,+0.096]	−
crpo-sft-llama-3.1-8b-instruct	high	+0.581	[+0.303,+0.789]	+	+0.375	[-0.009,+0.770]	+
crpo-sft-llama-3.1-8b-instruct	low	+0.403	[+0.071,+0.710]	+	+0.234	[-0.237,+0.678]	±
deepseek-r1-distill-llama-8b	high	-0.092	[-0.364,+0.172]	±	-0.172	[-0.388,-0.019]	−
deepseek-r1-distill-llama-8b	low	+0.166	[-0.248,+0.426]	±	-0.015	[-0.336,+0.290]	±
deepseek-r1-distill-qwen-7b	high	+0.338	[-0.014,+0.645]	+	+0.132	[-0.075,+0.460]	+
deepseek-r1-distill-qwen-7b	low	+0.244	[-0.026,+0.612]	+	+0.267	[+0.129,+0.442]	+
llama-3.1-8b	high	+0.003	[-0.332,+0.207]	±	+0.039	[-0.245,+0.310]	±
llama-3.1-8b	low	+0.424	[-0.089,+0.605]	+	-0.114	[-0.315,+0.258]	±
llama-3.1-8b-instruct	high	+0.113	[-0.099,+0.421]	+	+0.221	[+0.025,+0.549]	+
llama-3.1-8b-instruct	low	+0.326	[+0.157,+0.494]	+	+0.314	[+0.012,+0.547]	+
qwen2.5-7b	high	-0.042	[-0.316,+0.216]	±	-0.326	[-0.564,-0.079]	−
qwen2.5-7b	low	+0.079	[-0.178,+0.256]	±	+0.109	[-0.183,+0.406]	±
qwen2.5-7b-instruct	high	+0.086	[-0.215,+0.361]	±	+0.072	[-0.334,+0.359]	±
qwen2.5-7b-instruct	low	+0.145	[-0.121,+0.285]	±	+0.035	[-0.216,+0.268]	±
qwen2.5-math-7b	high	+0.051	[-0.225,+0.432]	±	+0.175	[-0.098,+0.444]	+
qwen2.5-math-7b	low	+0.182	[-0.072,+0.477]	+	+0.076	[-0.083,+0.331]	+

Table 2: Default Mode Network (DMN) AUT alignment results (prompt+response activations) per model and creativity population with rating cutoff=2.0.

model	pop	last-token			mean-token		
		median	95% CI	alignment	median	95% CI	alignment
crpo-dpo-llama-3.1-8b-instruct	high	+0.467	[-0.049,+1.000]	+	+0.091	[-0.282,+0.479]	±
crpo-dpo-llama-3.1-8b-instruct	low	+0.454	[+0.115,+0.862]	+	+0.253	[-0.184,+0.670]	±
crpo-llama-3.1-8b-instruct-cre	high	+0.366	[-0.228,+0.666]	±	+0.088	[-0.135,+0.629]	±
crpo-llama-3.1-8b-instruct-cre	low	+0.112	[-0.183,+0.360]	±	-0.040	[-0.252,+0.077]	−
crpo-sft-llama-3.1-8b-instruct	high	+0.633	[+0.324,+0.848]	+	+0.454	[-0.173,+0.769]	±
crpo-sft-llama-3.1-8b-instruct	low	+0.331	[+0.170,+0.485]	+	-0.126	[-0.506,+0.279]	±
deepseek-r1-distill-llama-8b	high	-0.486	[-0.833,-0.320]	−	-0.248	[-0.524,+0.024]	−
deepseek-r1-distill-llama-8b	low	+0.297	[+0.124,+0.550]	+	+0.100	[-0.246,+0.301]	±
deepseek-r1-distill-qwen-7b	high	+0.263	[-0.206,+0.835]	±	-0.197	[-0.520,+0.157]	±
deepseek-r1-distill-qwen-7b	low	+0.161	[-0.082,+0.527]	+	-0.020	[-0.281,+0.428]	±
llama-3.1-8b	high	+0.088	[-0.238,+0.470]	±	+0.040	[-0.500,+0.318]	±
llama-3.1-8b	low	+0.452	[+0.146,+1.000]	+	+0.019	[-0.428,+0.268]	±
llama-3.1-8b-instruct	high	+0.495	[-0.041,+0.862]	+	+0.094	[-0.324,+0.503]	±
llama-3.1-8b-instruct	low	+0.289	[+0.155,+0.595]	+	+0.197	[-0.020,+0.472]	+
qwen2.5-7b	high	+0.033	[-0.394,+0.234]	±	+0.044	[-0.338,+0.558]	±
qwen2.5-7b	low	+0.034	[-0.206,+0.309]	±	-0.016	[-0.189,+0.163]	±
qwen2.5-7b-instruct	high	+0.068	[-0.440,+0.369]	±	+0.351	[-0.112,+0.748]	±
qwen2.5-7b-instruct	low	-0.034	[-0.308,+0.254]	±	+0.061	[-0.186,+0.266]	±
qwen2.5-math-7b	high	-0.197	[-0.572,+0.179]	±	-0.258	[-0.657,+0.033]	−
qwen2.5-math-7b	low	+0.404	[+0.086,+0.755]	+	+0.268	[-0.049,+0.432]	+

Table 3: Default Mode Network (DMN) AUT alignment results (prompt+response activations) per model and creativity population with rating cutoff=2.25.

fold at that layer. Every subject thus receives a score from a layer selected without them. We repeat the partition 20 times and average. Selection and evaluation never share subjects, so the maximum operation cannot inflate the evaluated value.

The bias is substantial but does not scale with layer count. Cross-validation reduces alignment scores by a median of 0.13 (median 33% of the naive value; range 0.007–0.267), confirming that best-layer scores are upper bounds rather than unbiased estimates. However, the reduction is *negatively* related to layer count ($r = -0.48$, $p = 0.018$), the opposite of what the confound requires. What it tracks instead is how much genuine layer structure a model

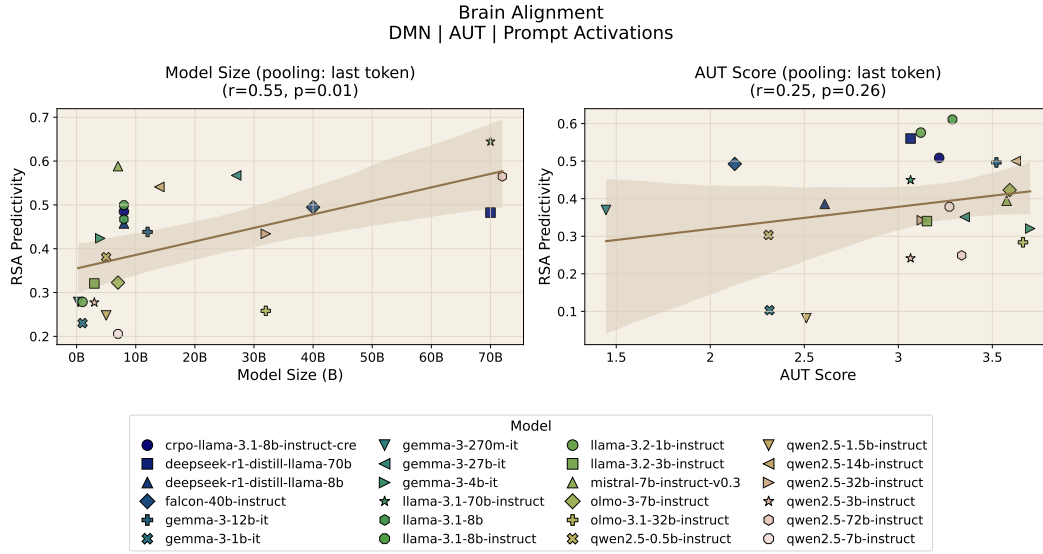


Figure 12: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using prompt-only model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

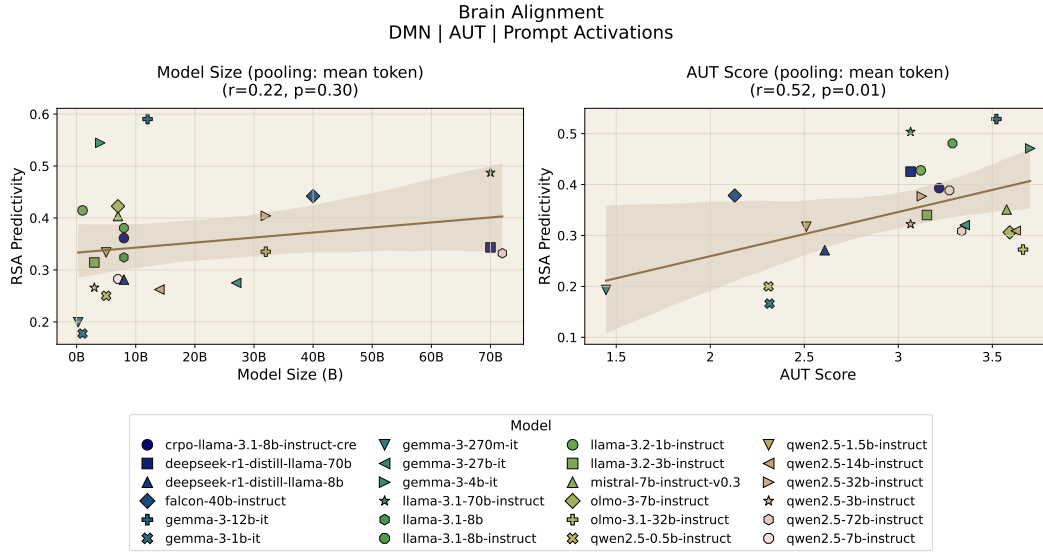


Figure 13: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using prompt-only model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

has: the gap is strongly negatively related to cross-validated alignment ($r = -0.81$, $p < 0.001$). Models with a reliable alignment peak lose almost nothing under cross-validation (QWEN2.5-72B-INSTRUCT: $0.565 \rightarrow 0.557$), whereas models without one lose most of their apparent alignment (GEMMA-3-270M-IT: $0.278 \rightarrow 0.011$), because in the latter case the maximum reflects noise rather than signal.

The size correlation survives. Because the bias is larger for weakly-aligned models, which are predominantly the smaller ones, it attenuates rather than manufactures a positive size relationship. Recomputing the DMN-size correlation on cross-validated alignment values

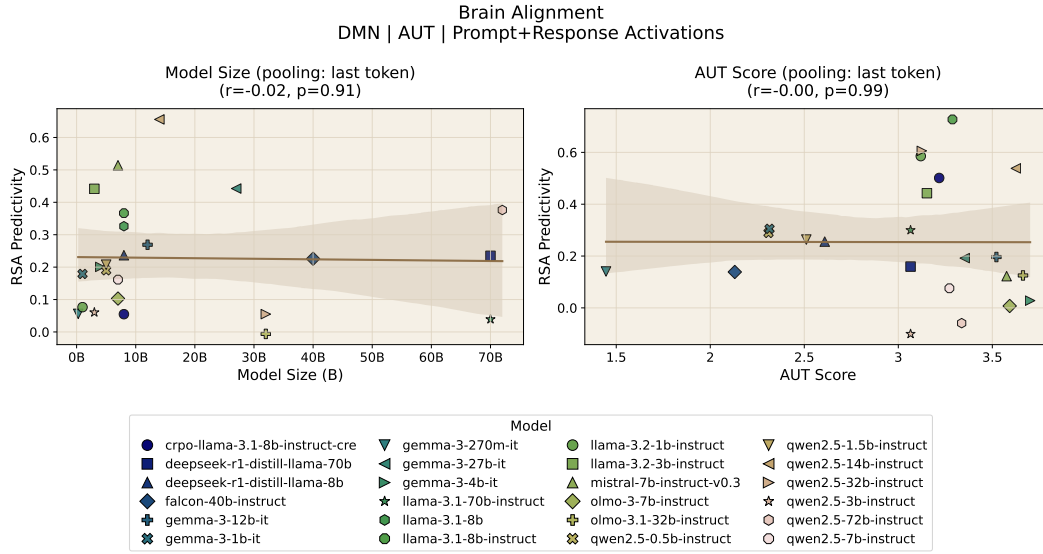


Figure 14: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

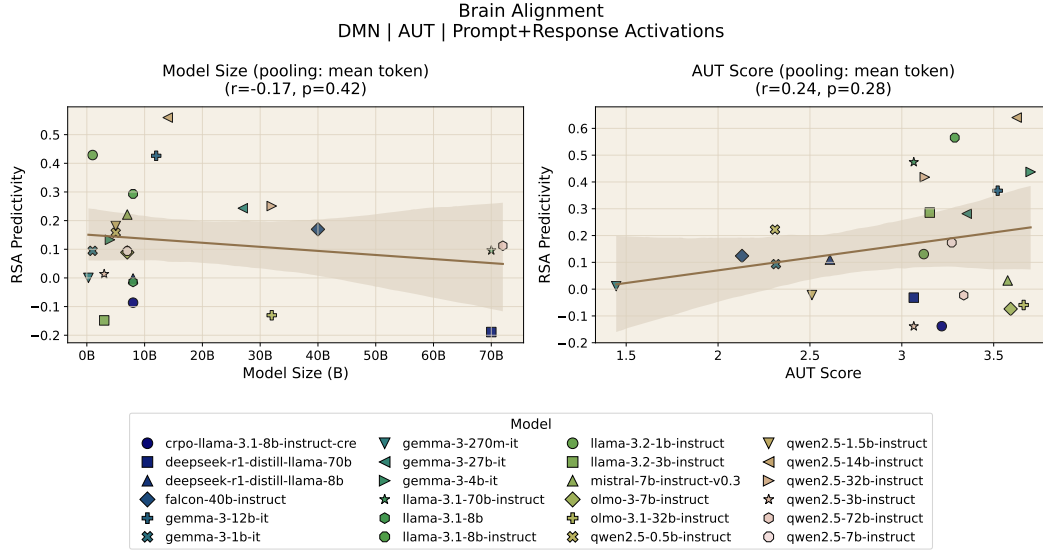


Figure 15: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

leaves it essentially unchanged, and marginally stronger: $r = +0.56$ ($p = 0.004$) versus $r = +0.55$ ($p = 0.006$) under the naive estimator. We conclude that this correlation is not an artifact of layer selection.

Modified AUT Instructions for scoring

Think of creative uses for objects. Report only the most original idea in a few words.

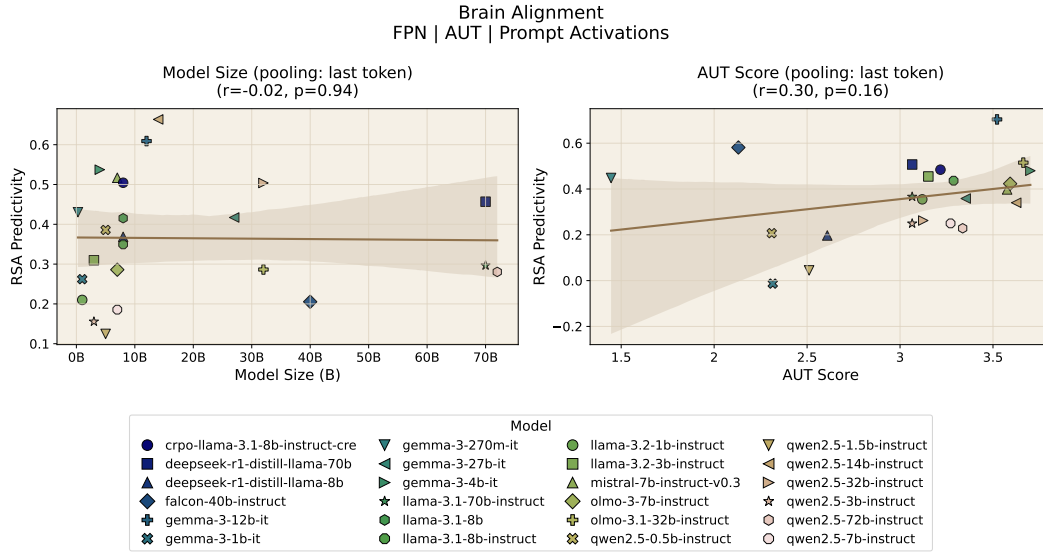


Figure 16: **Frontoparietal Network (FPN) AUT brain alignment results by model size and task performance using prompt-only model activations with last-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

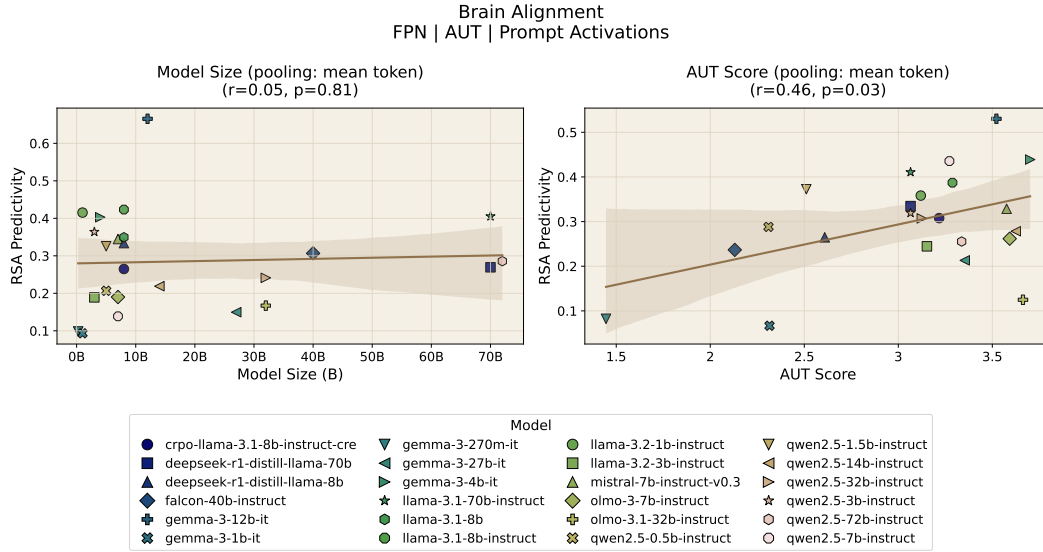


Figure 17: **Frontoparietal Network (FPN) AUT brain alignment results by model size and task performance using prompt-only model activations with mean-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

Modified OCT Instructions for scoring

Think of the physical properties of objects. Report only the most prominent characteristic in a few words.

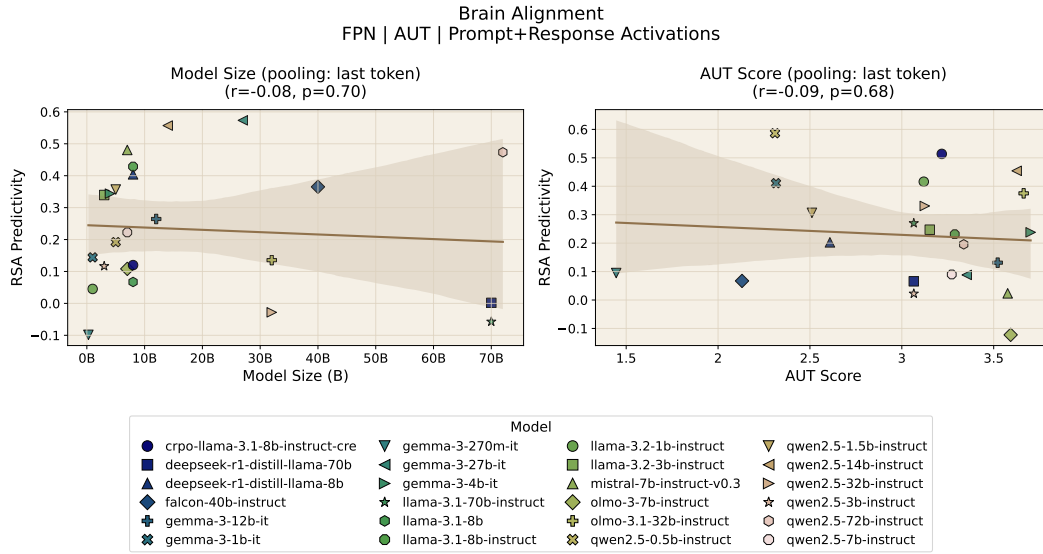


Figure 18: Frontoparietal Network (FPN) AUT brain alignment results by model size and task performance using prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

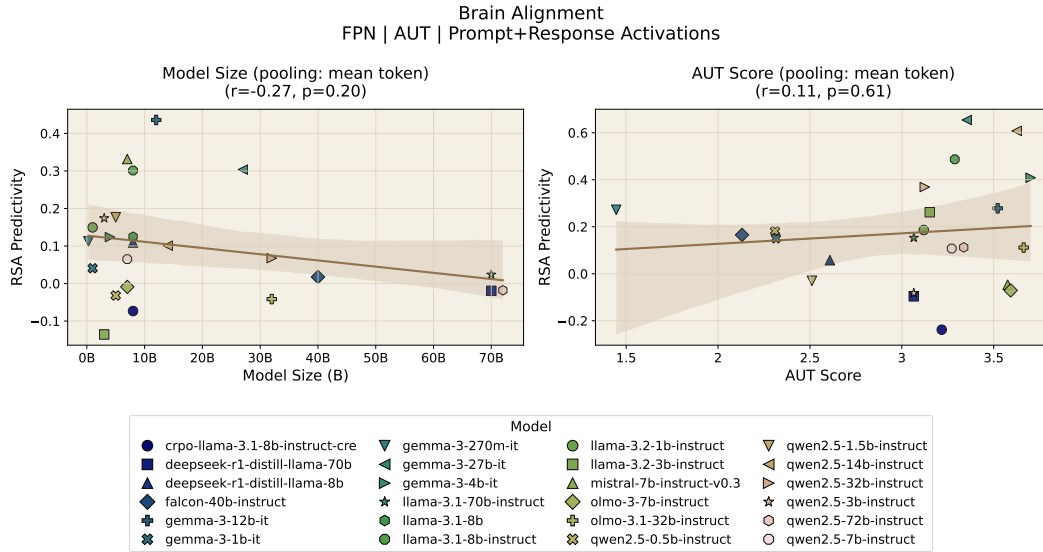


Figure 19: Frontoparietal Network (FPN) AUT brain alignment results by model size and task performance using prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

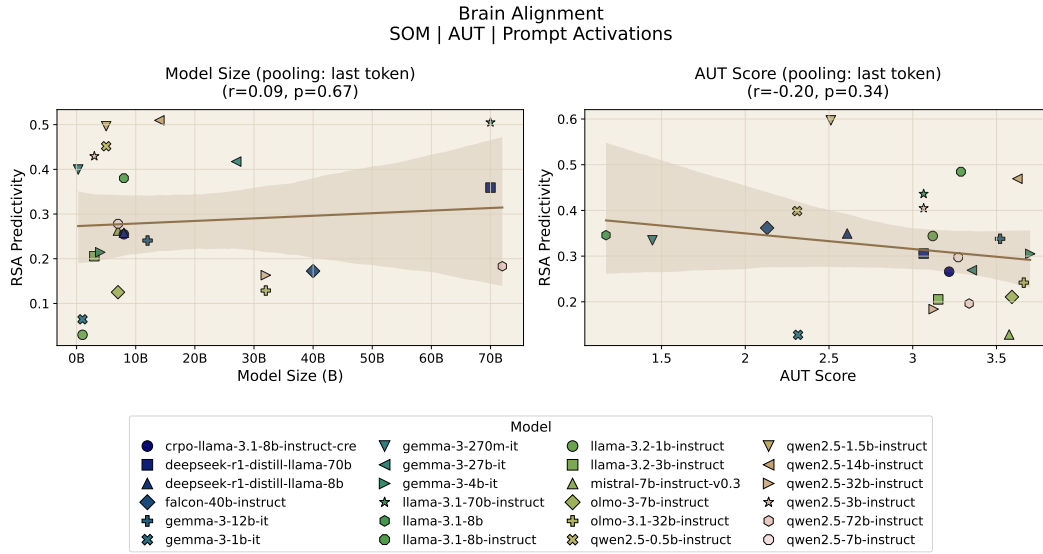


Figure 20: **Somatomotor Network (SOM) AUT brain alignment results by model size and task performance using prompt-only model activations with last-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

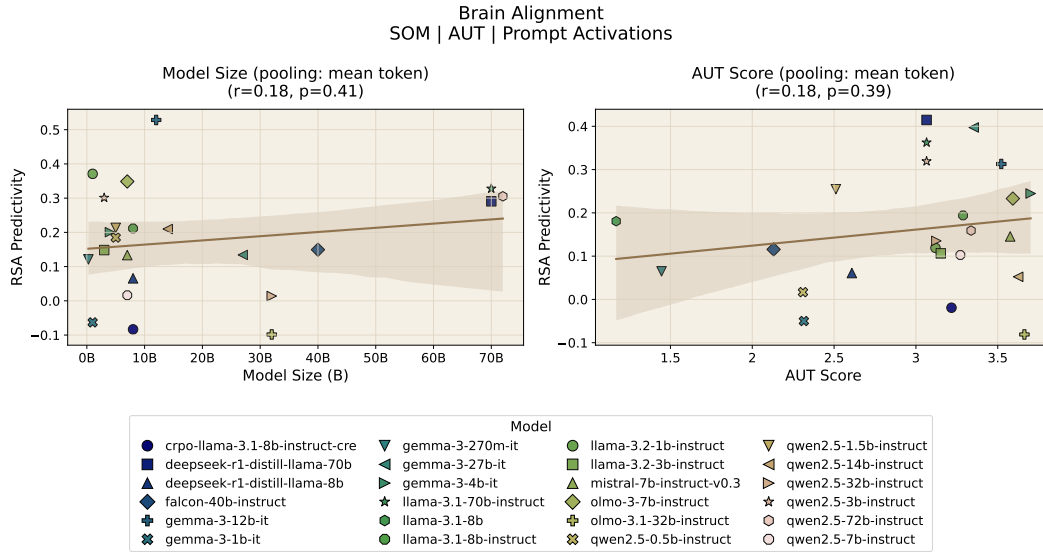


Figure 21: **Somatomotor Network (SOM) AUT brain alignment results by model size and task performance using prompt-only model activations with mean-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

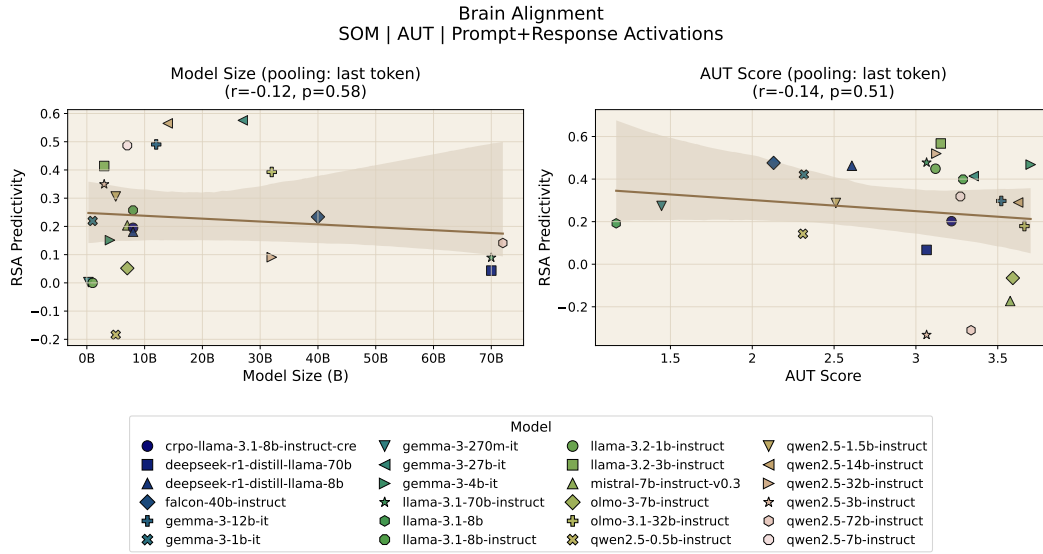


Figure 22: Somatomotor Network (SOM) AUT brain alignment results by model size and task performance using prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

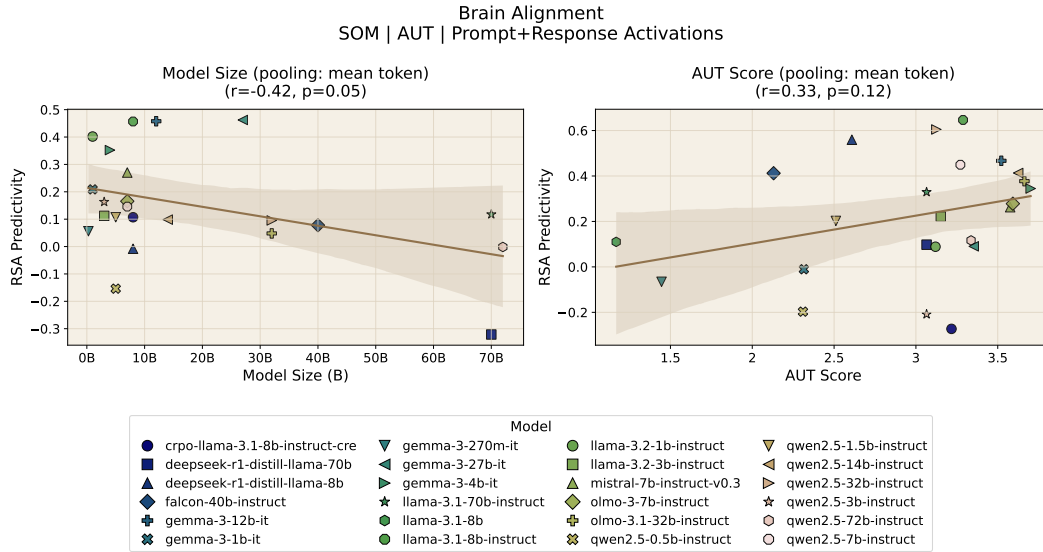


Figure 23: Somatomotor Network (SOM) AUT brain alignment results by model size and task performance using prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

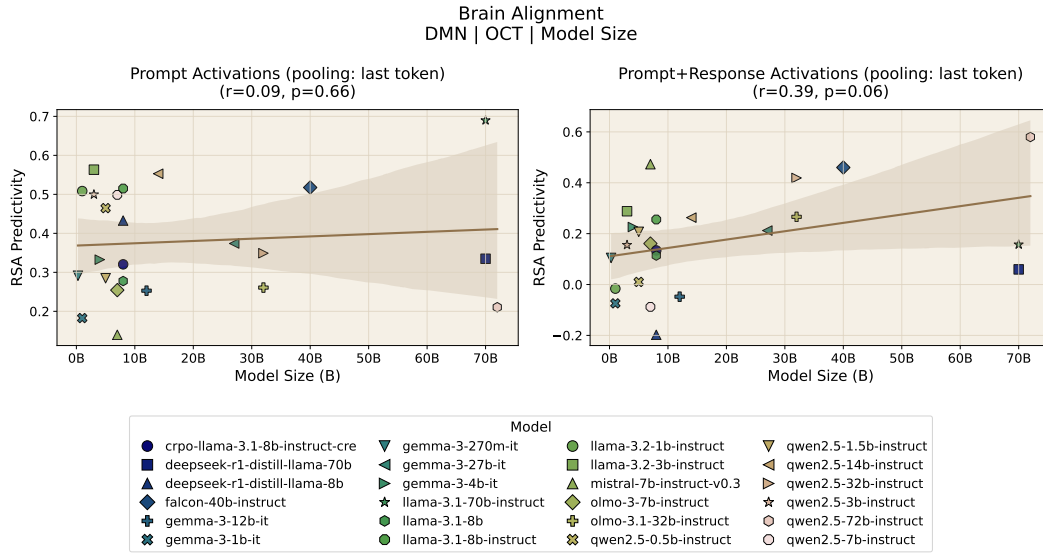


Figure 24: Default Mode Network (DMN) OCT brain alignment results by model size with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

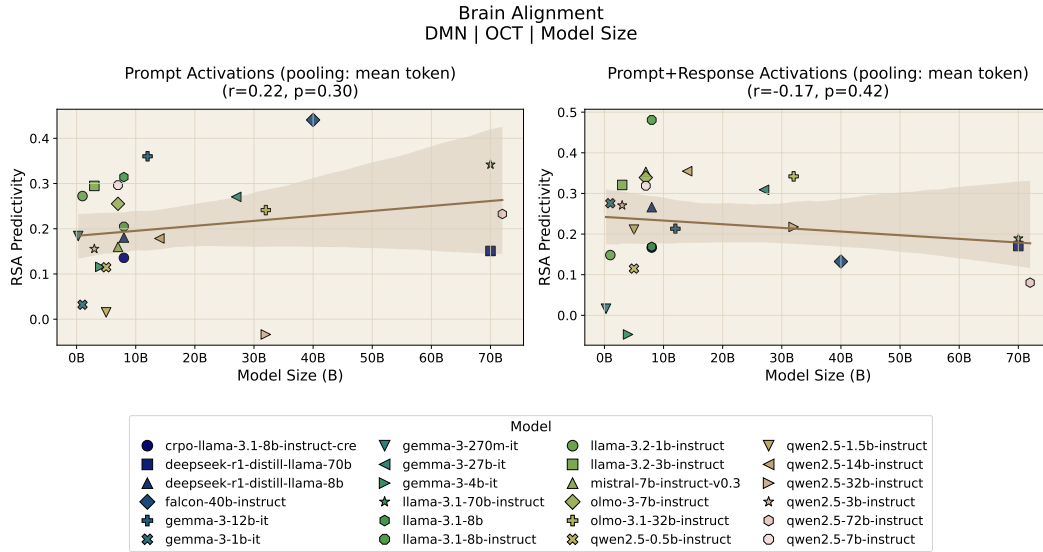


Figure 25: Default Mode Network (DMN) OCT brain alignment results by model size with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

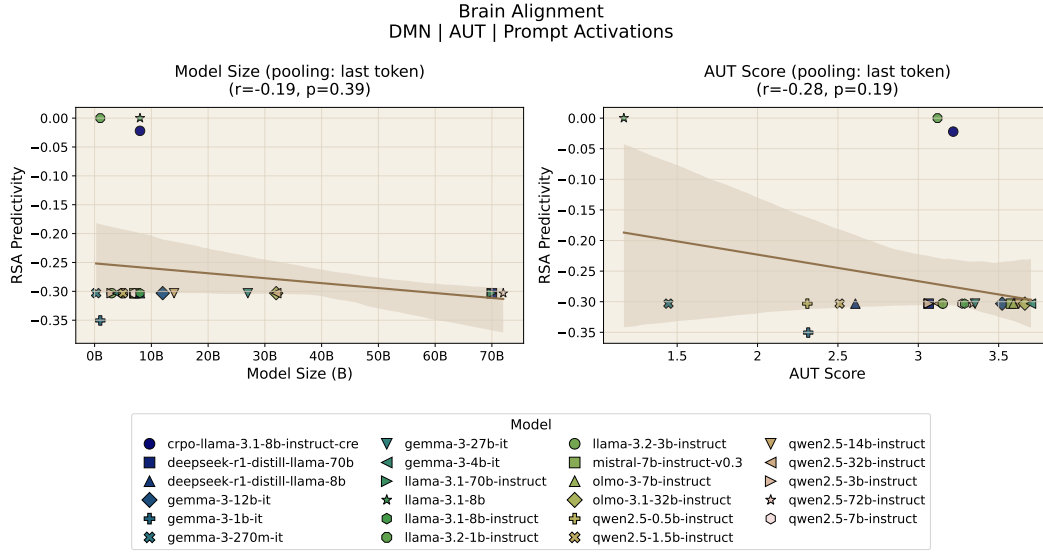


Figure 26: **Default Mode Network (DMN) AUT brain alignment results by model size and task performance using empty prompt-only model activations with last-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value. Note that FALCON-40B-INSTRUCT result is missing due to the failure to extract representations from an empty prompt.

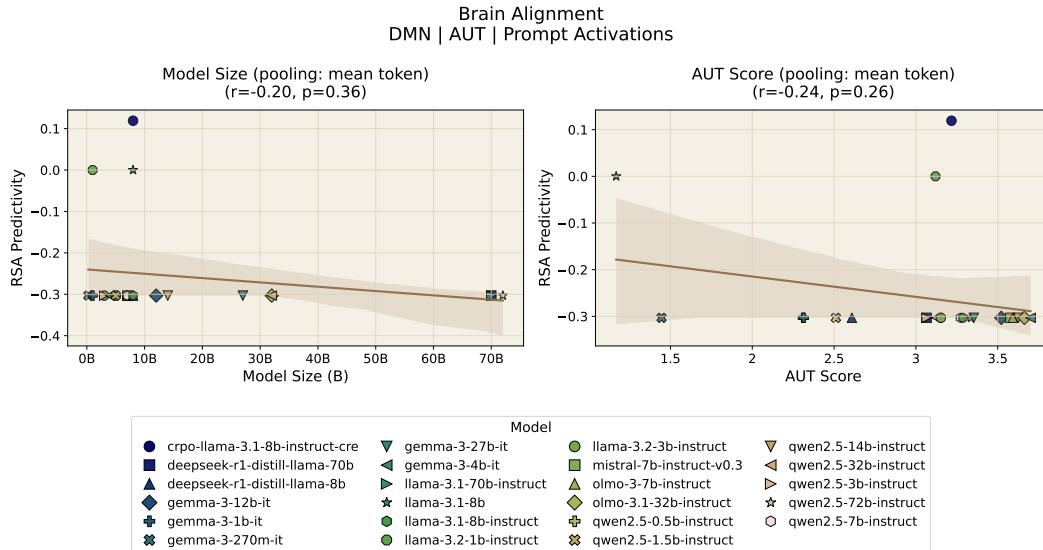


Figure 27: **Default Mode Network (DMN) AUT brain alignment results by model size and task performance using empty prompt-only model activations with mean-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value. Note that FALCON-40B-INSTRUCT result is missing due to the failure to extract representations from an empty prompt.

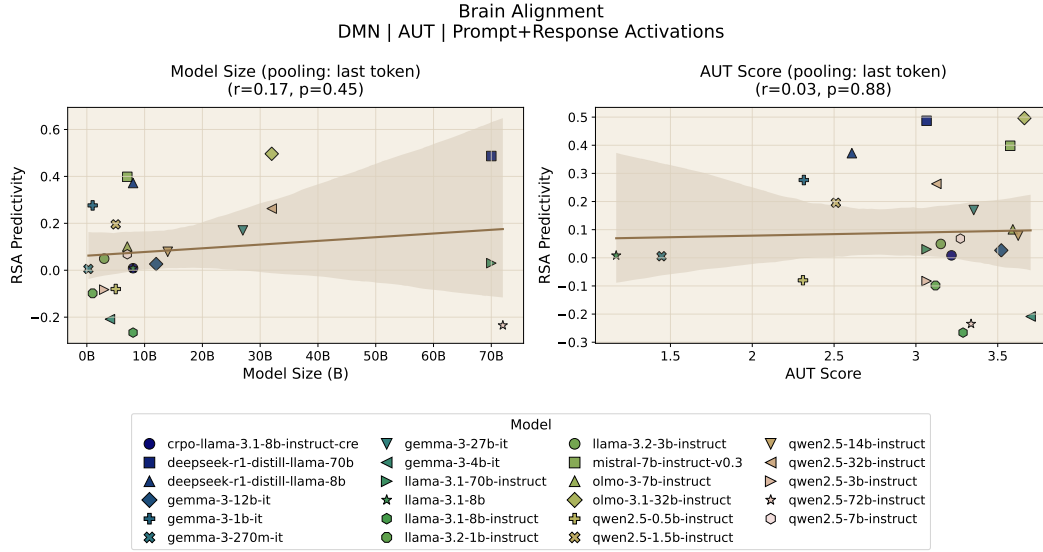


Figure 28: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using empty prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value. Note that FALCON-40B-INSTRUCT result is missing due to the failure to extract representations from an empty prompt.

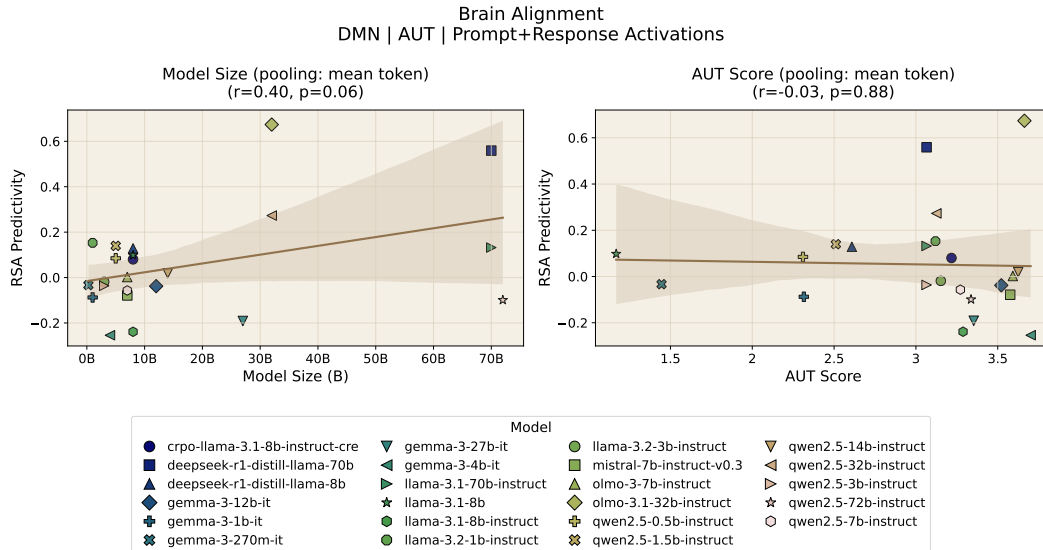


Figure 29: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using empty prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value. Note that FALCON-40B-INSTRUCT result is missing due to the failure to extract representations from an empty prompt.

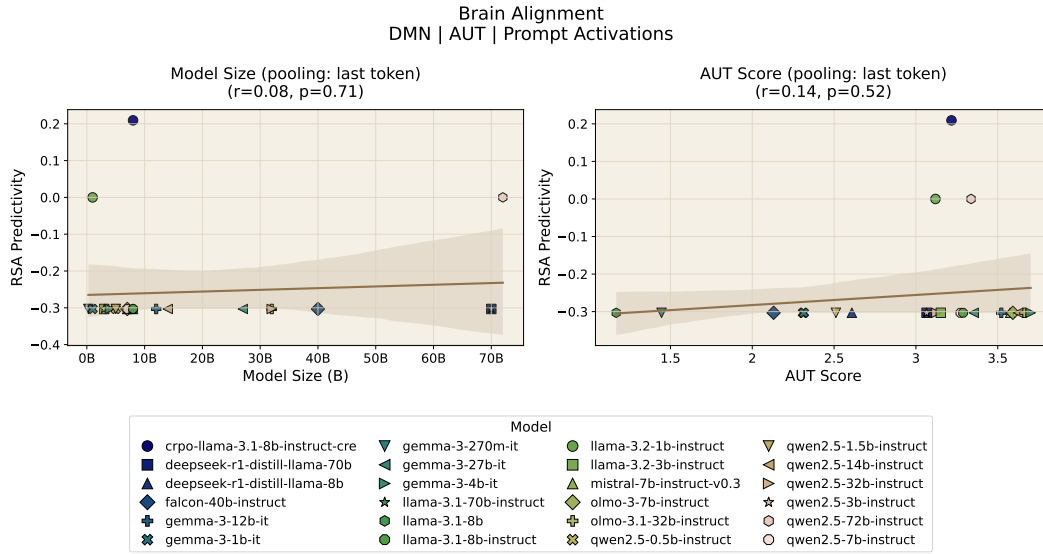


Figure 30: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using placeholder prompt-only model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

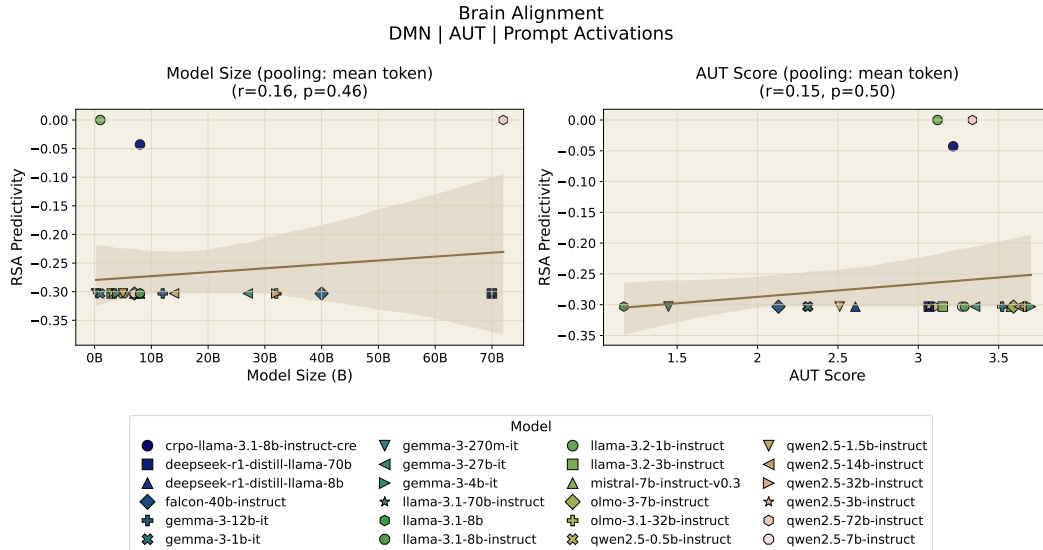


Figure 31: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using placeholder prompt-only model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

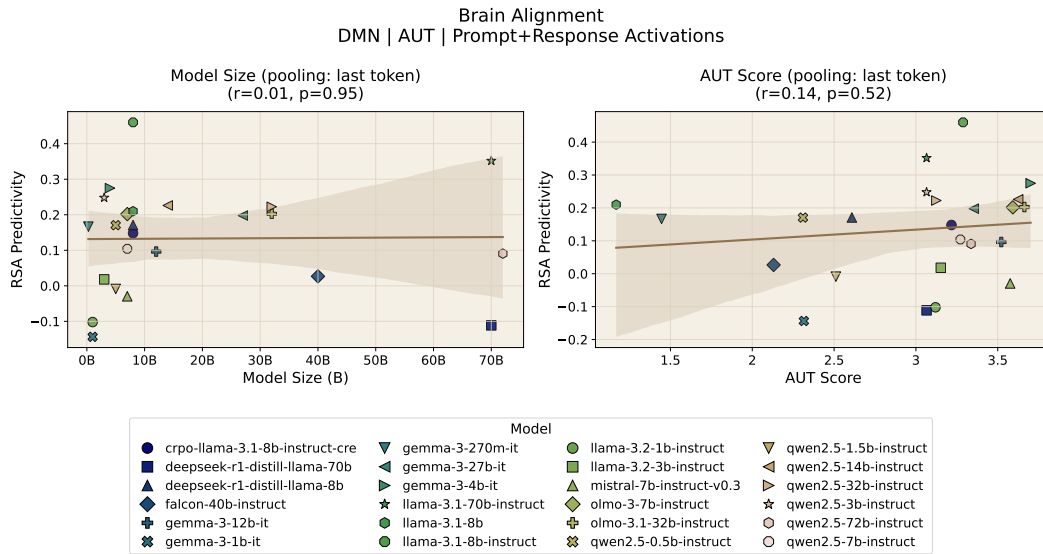


Figure 32: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using placeholder prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

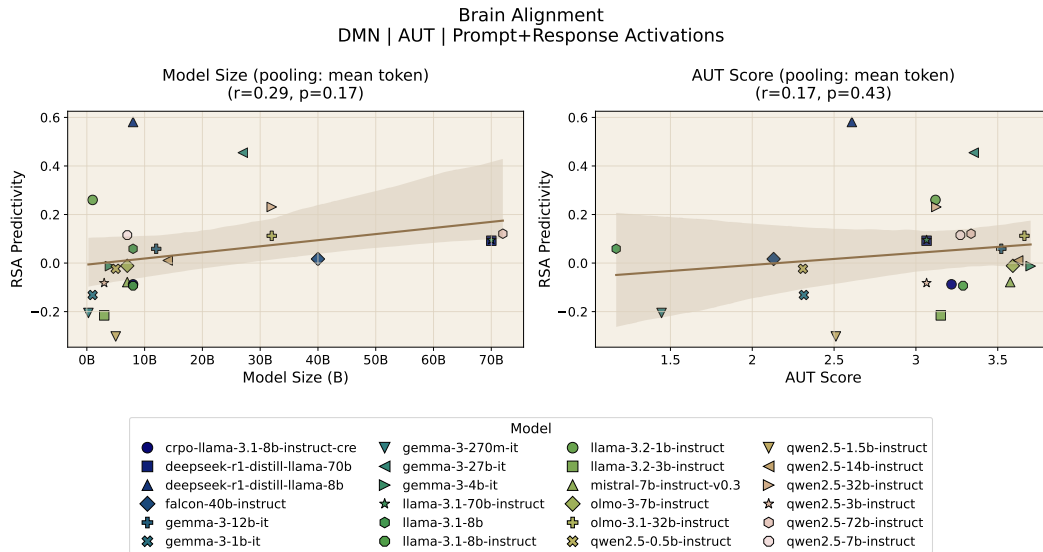


Figure 33: Default Mode Network (DMN) AUT brain alignment results by model size and task performance using placeholder prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

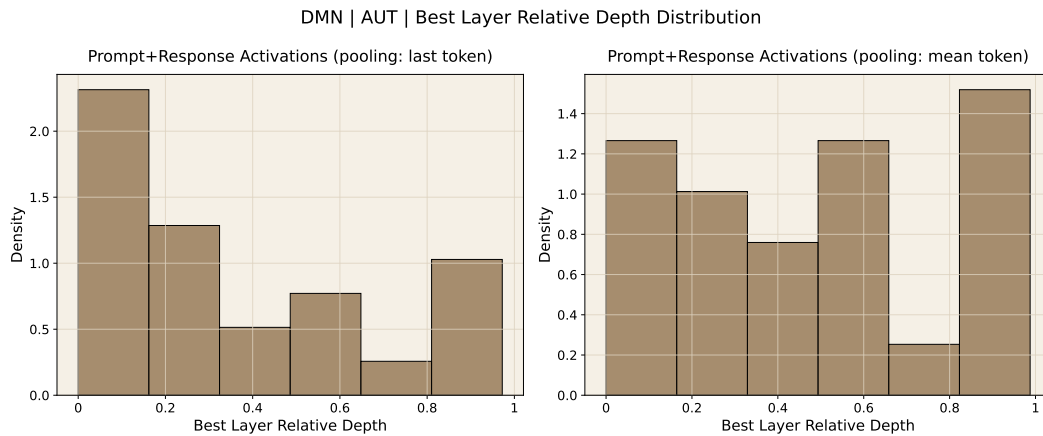


Figure 34: The distribution of best model layer (measured by relative depth) for DMN AUT brain alignment on prompt+response model activations.

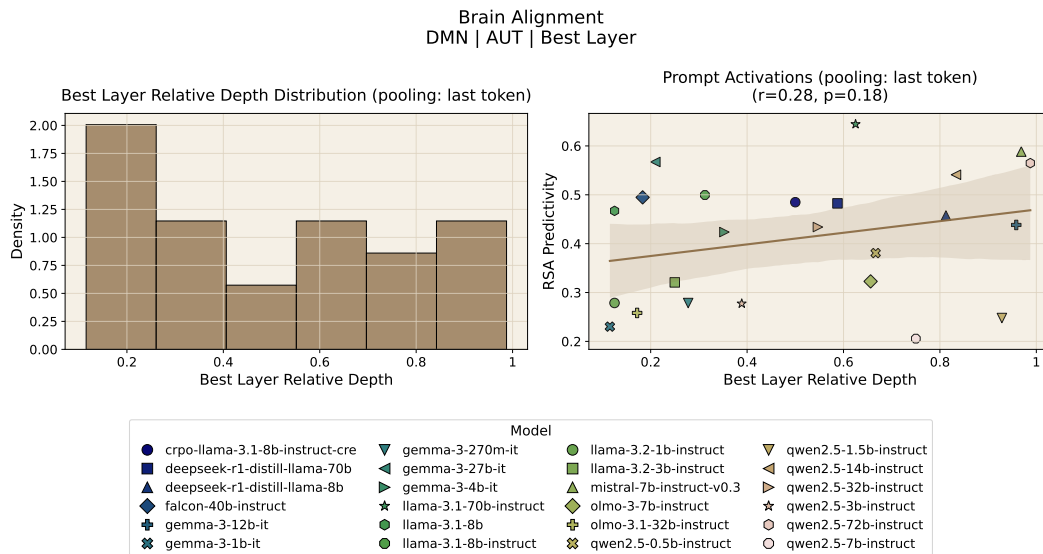


Figure 35: The distribution of best model layer (measured by relative depth) and its correlation with DMN AUT brain alignment on prompt-only model activations with last-token pooling.

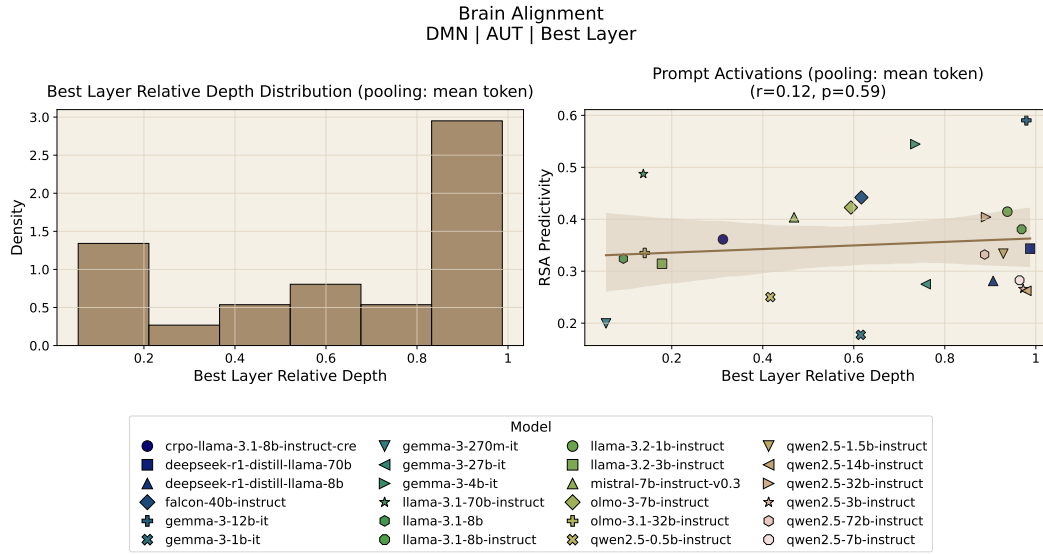


Figure 36: The distribution of best model layer (measured by relative depth) and its correlation with DMN AUT brain alignment on prompt-only model activations with mean-token pooling.

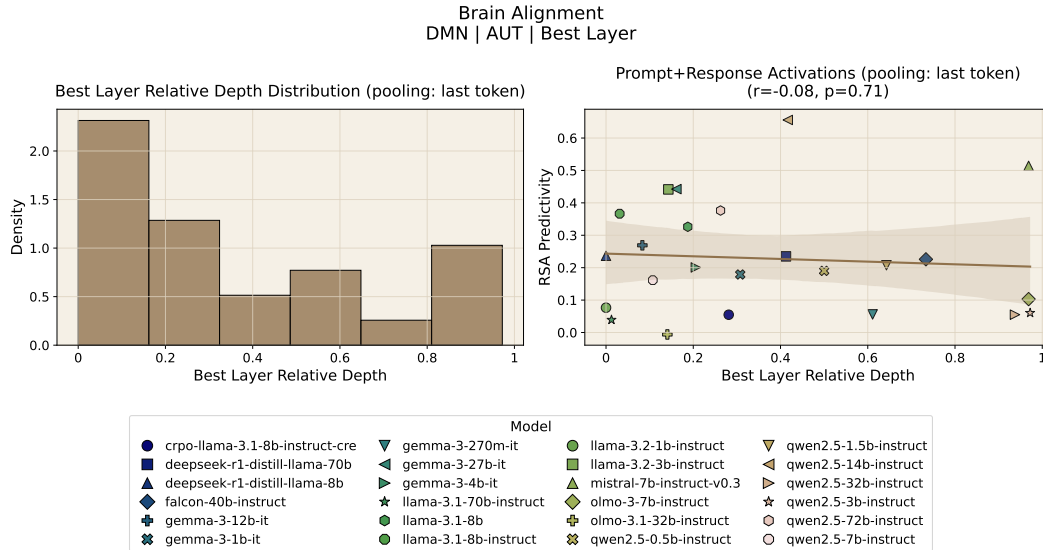


Figure 37: The distribution of best model layer (measured by relative depth) and its correlation with DMN AUT brain alignment on prompt+response model activations with last-token pooling.

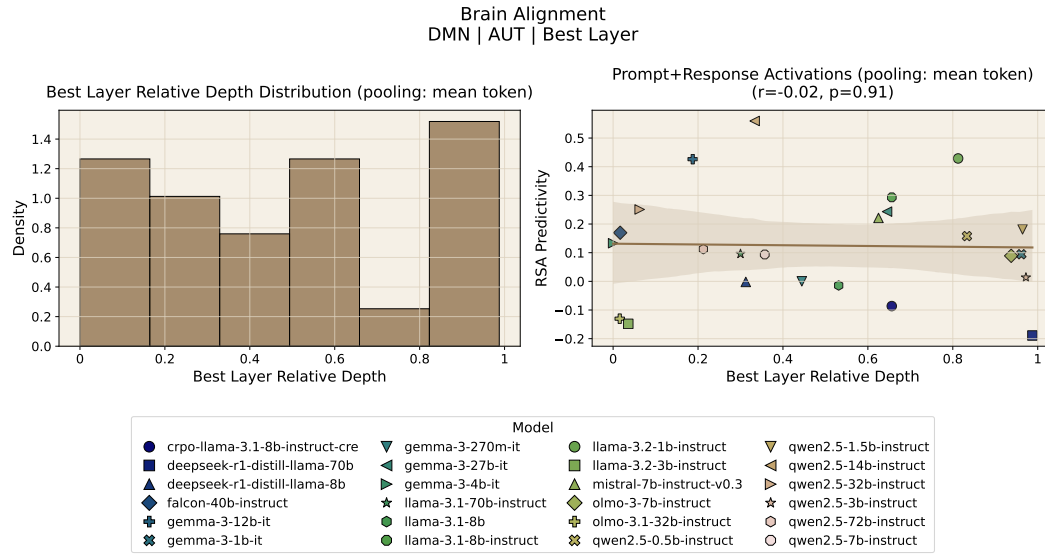


Figure 38: The distribution of best model layer (measured by relative depth) and its correlation with DMN AUT brain alignment on prompt+response model activations with mean-token pooling.

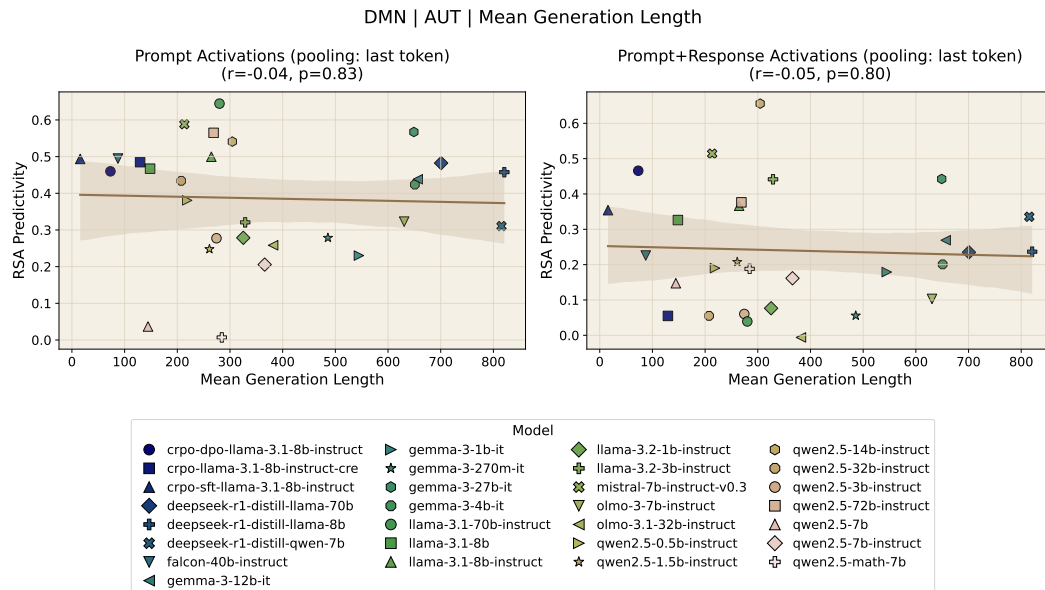


Figure 39: Default Mode Network (DMN) AUT brain alignment results by mean response generation length using prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

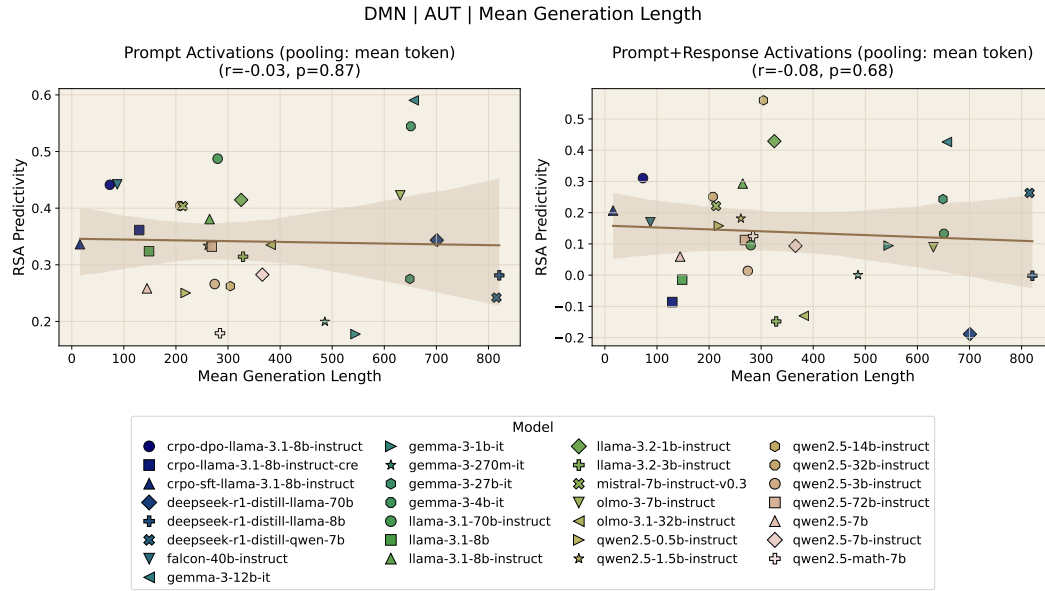


Figure 40: Default Mode Network (DMN) AUT brain alignment results by mean response generation length using prompt+response model activations with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

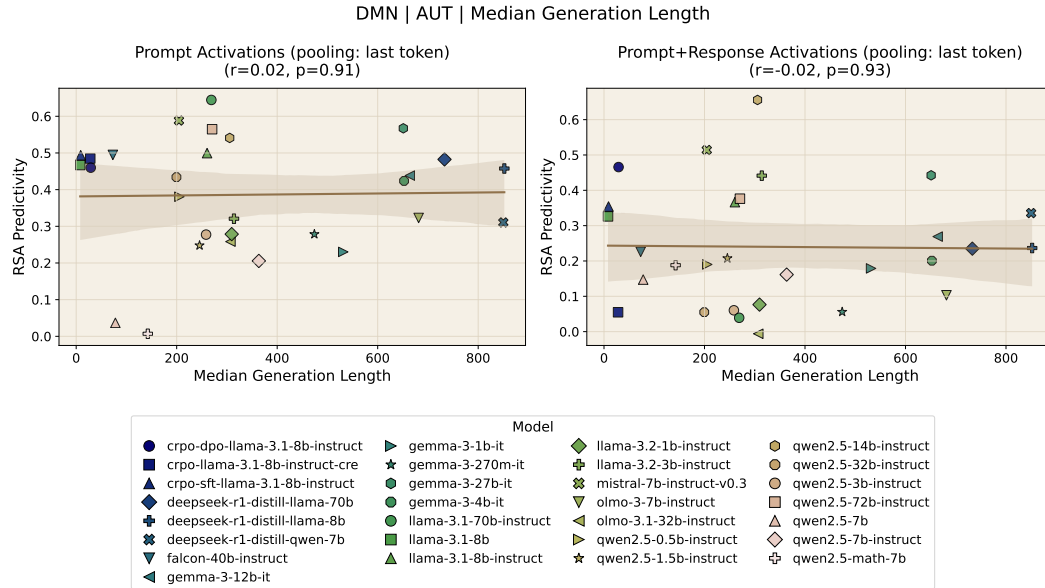


Figure 41: Default Mode Network (DMN) AUT brain alignment results by median response generation length using prompt+response model activations with last-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

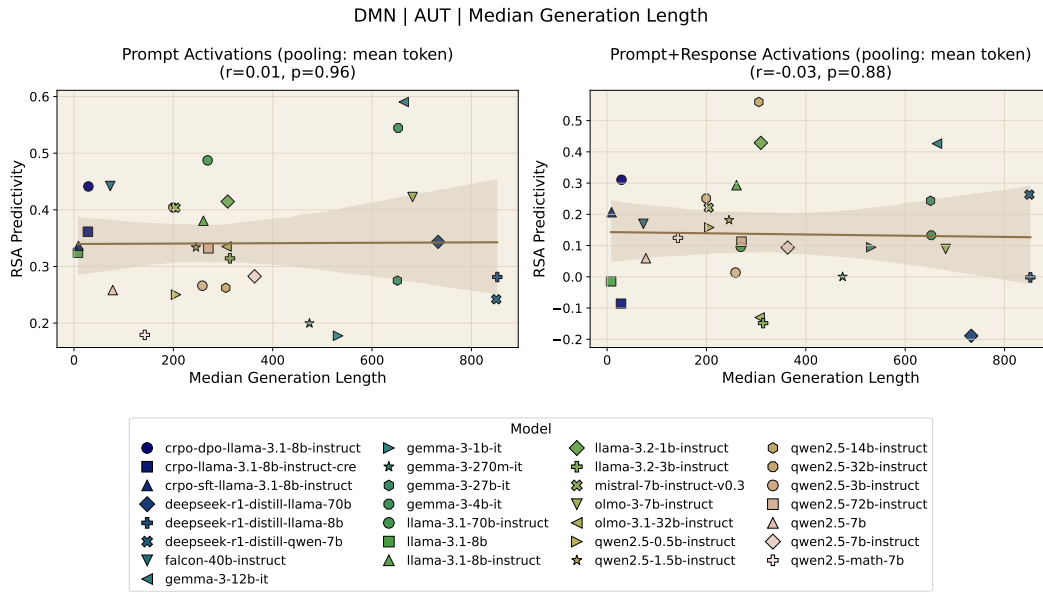


Figure 42: **Default Mode Network (DMN) AUT brain alignment results by median response generation length using prompt+response model activations with mean-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

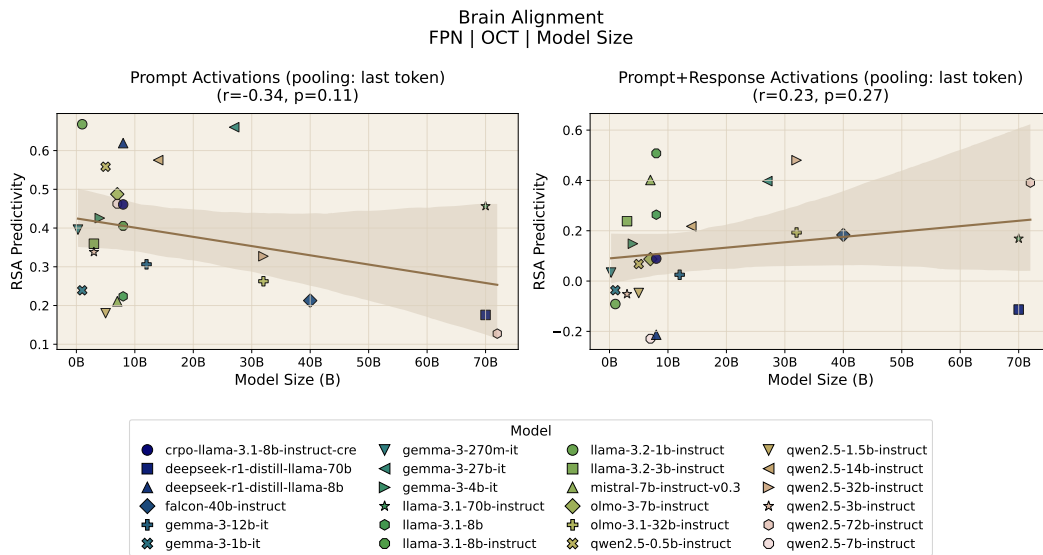


Figure 43: **Frontoparietal network (FPN) OCT brain alignment results by model size with last-token pooling.** r and p correspond to the Pearson correlation coefficient and p-value.

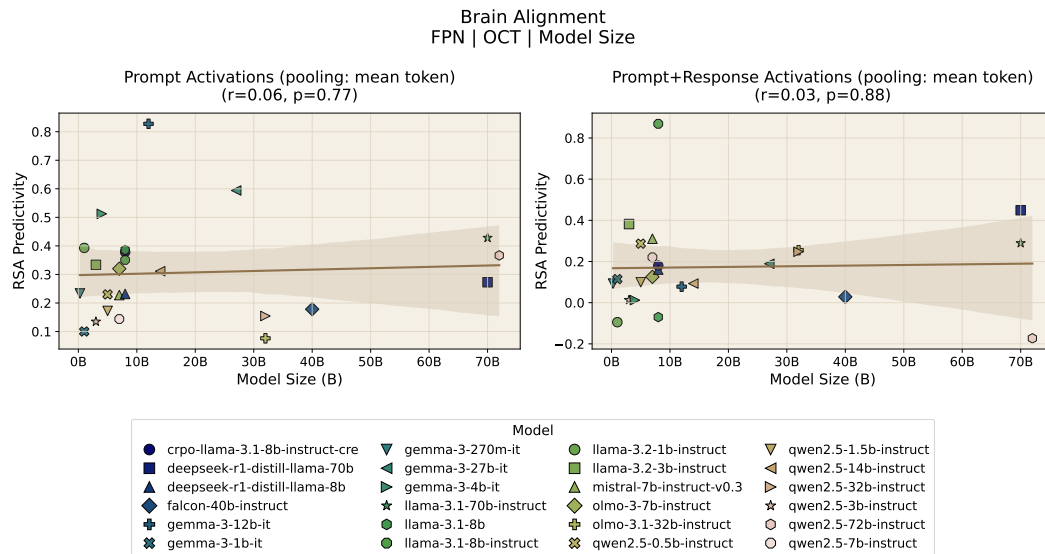


Figure 44: Frontoparietal network (FPN) OCT brain alignment results by model size with mean-token pooling. r and p correspond to the Pearson correlation coefficient and p-value.

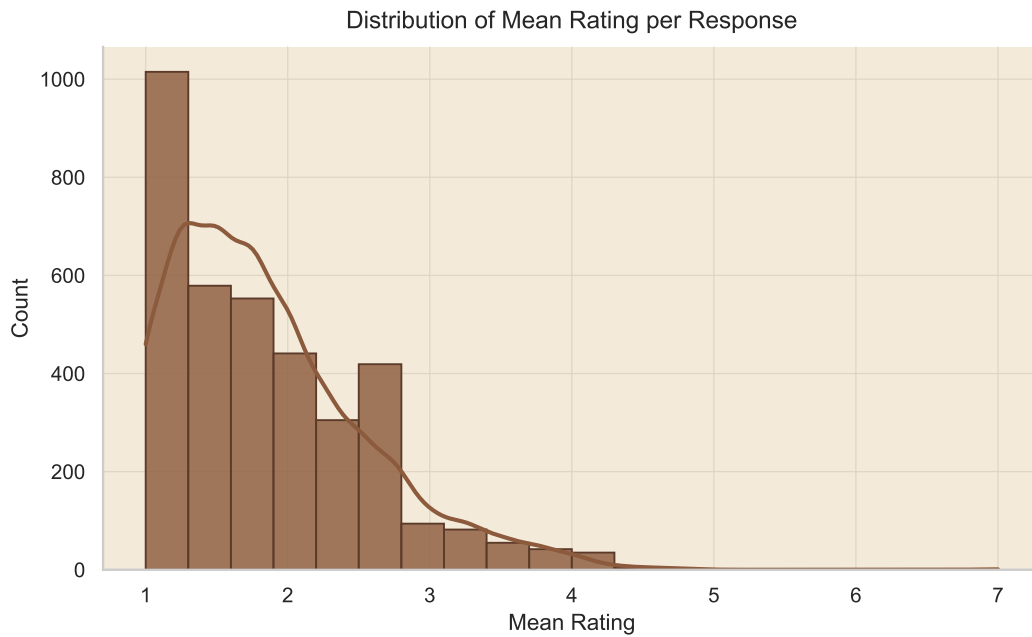


Figure 45: Distribution of mean AUT ratings per human response.

Model	Mean	Median	Std
llama-3.2-1b-instruct	325.27	309.50	87.26
llama-3.2-3b-instruct	328.85	314.00	91.82
llama-3.1-8b	148.34	8.00	292.45
llama-3.1-8b-instruct	264.61	260.50	61.98
crpo-llama-3.1-8b-instruct-cre	129.39	28.00	273.30
crpo-sft-llama-3.1-8b-instruct	15.71	9.00	65.67
crpo-dpo-llama-3.1-8b-instruct	73.09	29.00	178.99
deepseek-r1-distill-llama-8b	820.72	852.00	85.06
deepseek-r1-distill-llama-70b	700.77	733.00	149.62
llama-3.1-70b-instruct	280.10	269.00	81.04
gemma-3-270m-it	485.91	474.00	127.26
gemma-3-1b-it	544.76	531.50	78.84
gemma-3-4b-it	650.92	652.50	52.31
gemma-3-12b-it	656.64	664.00	40.42
gemma-3-27b-it	649.09	651.00	27.51
olmo-3-7b-instruct	630.65	681.50	153.70
olmo-3.1-32b-instruct	381.91	307.00	203.53
falcon-40b-instruct	87.33	73.00	75.00
qwen2.5-0.5b-instruct	218.70	205.50	85.65
qwen2.5-1.5b-instruct	261.00	245.50	120.88
qwen2.5-3b-instruct	274.38	258.50	98.06
deepseek-r1-distill-qwen-7b	815.35	850.00	100.00
qwen2.5-math-7b	284.51	142.50	281.23
qwen2.5-7b	144.43	78.00	161.71
qwen2.5-7b-instruct	365.88	363.50	85.31
qwen2.5-14b-instruct	304.46	305.50	56.72
qwen2.5-32b-instruct	207.46	199.50	58.42
qwen2.5-72b-instruct	268.90	270.50	100.12
mistral-7b-instruct-v0.3	213.41	204.50	54.92

Table 4: Generation length (in words) statistics per model over the AUT responses.