WHICH ATTENTION HEADS ARE LIKE THE HUMAN HEAD? NOT THE ONES THAT COMPUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain–AI alignment is often interpreted as a sign that model and brain perform similar computations. Whether the aligned units are causally involved in model computation is rarely checked. On an abstract pattern-completion task (AAABAAA → B), we compare LLM attention-head representations with human EEG and test how ablating those heads affects task performance. Alignment and causation dissociate: brain-aligned heads contribute to performance, but their removal is substantially less disruptive than removal of heads selected via attribution patching. We compare two head sets that prior interpretability work defines without reference to the brain: concept vectors (CVs), which represent abstract patterns across formats, and function vectors (FVs), selected for their contribution to correct-answer prediction. Brain alignment shows little association with FV scores, while its association with CV scores varies across models. Among brain-aligned heads, we find recurring attention profiles: one emphasizes distinctive elements (novelty heads), the other repeating elements (repetition heads). The novelty family tracks salience and attends to the same elements that humans look at, yet its removal is less damaging than random ablation on average. Repetition heads contribute modestly to performance and are associated with abstract-pattern representation (CVs). Across 17 models spanning 3B–72B parameters, FV-ranked removal is substantially more disruptive than brain-ranked removal. Brain alignment thus captures how the model reads the stimulus, and only faintly captures how it represents the pattern and solves the task.

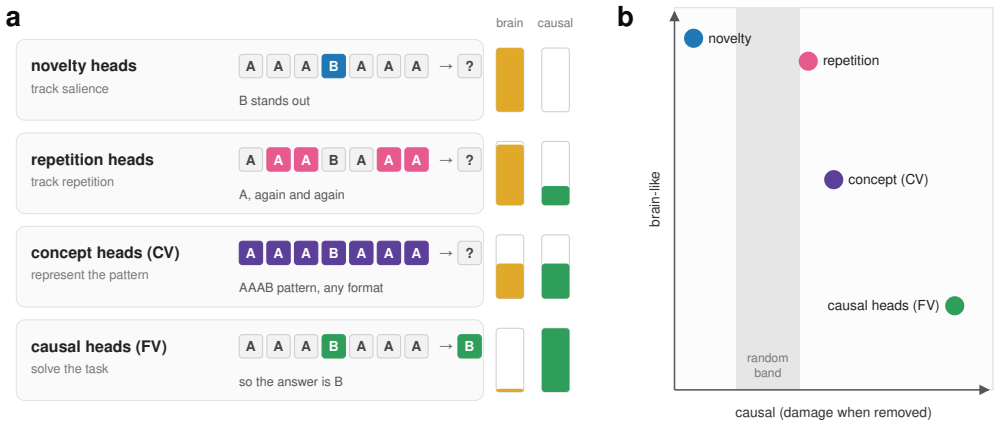

Figure 1: **Schematic description of the findings.** (a) What each attention head population does with one item of the task, the pattern AAABAAA whose answer is B (in words, for example, "hat hat hat risk hat hat hat" with answer "risk"), beside its brain alignment and its causal contribution. Colored symbols mark what each population attends to or encodes. (b) The same populations, schematically, in the plane of brain alignment against damage when the heads are removed; the gray band marks the damage from ablating random heads. Measured ablation curves are in Figure 3.

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

1 INTRODUCTION

When a unit inside an artificial neural network predicts a neural signal recorded from a person performing the same task, the result is reported as *alignment* and often read as evidence that the two systems compute something in common (Yamins et al., 2014; Schrimpf et al., 2018; 2021; Caucheteux & King, 2022; Goldstein et al., 2022). However, that reading has been challenged on logical and empirical grounds (Guest & Martin, 2023; Bowers et al., 2023; Antonello & Huth, 2024). The inference aspect has been tested indirectly: training a model to be brain-misaligned degrades its downstream performance (Merlin & Toneva, 2026), and units picked out by a neuroscience-style localizer are both causal and brain-aligned (AlKhamissi et al., 2025). What has not been done is the direct test: select units by their alignment with a neural signal recorded on a task, remove them, and measure the model's performance on that same task. In large language models (LLMs), attention heads provide a tractable unit for this test: their representations can be compared with neural recordings, and their contribution to task performance can be assessed by removing or patching their outputs. They also bring a prior that neural data alone cannot supply: mechanistic interpretability has already identified head classes with known roles, such as induction and copy heads (Olsson et al., 2022) and the function-vector heads of in-context learning (Todd et al., 2024), so a brain-alignment score can be read against heads whose function is known rather than against anonymous units.

We ask whether attention heads whose representations resemble human brain activity are also those on which the model's task performance most depends. We investigate this question using the abstract-sequence-completion task of Pinier et al. (2025): participants infer the missing eighth element from a sequence of icons arranged as one of eight distinct abstract patterns, such as `AAABAAA` $\rightarrow$ `B`. Models receive text-based versions, with words or symbols replacing the icons (Figure 2). Using fixation-related potentials (FRPs), EEG signals time-locked to eye-fixation onset, Pinier et al. (2025) reported positive representational correlations between frontal activity and LLM layers selected for their encoding of abstract patterns. We extend this layer-level comparison to individual attention heads and test its relationship to performance.

In LLMs, a dissociation between conceptual encoding (concept vectors; CVs) and representations that causally support task performance (function vectors; FVs) has been highlighted by Opiełka et al. (2026), building on the function-vector framework of Todd et al. (2024). On verbal analogies, heads selected for format-invariant concept representation generalized across formats and languages, whereas heads selected by activation patching supported in-context performance despite less consistent representations across formats. This distinction motivates two complementary analyses of the sequence-compltetion task. First, we score heads for concept representation and contribution to performance, independently of human data, and compare these scores with brain alignment. Second, we select the most brain-aligned heads, characterize their attention profiles, and test the effects of removing them. These analyses ask whether brain resemblance is more closely associated with abstract-pattern representation, with heads that matter most for predicting the answer or neither.

Our findings are as follows:

1. **Brain alignment does not track causal importance** (Section 2). Brain scores are weakly associated with attribution patching effects, and across models, removing brain-ranked heads is far less damaging than removing FV-ranked heads, the most causally important set.
2. **Brain-aligned heads reveal repetition and novelty profiles** (Section 3). Across models, two attention profiles consistently appear: one attends to repeated elements (the As in `AAABAAA`), the other to the distinctive element (the lone B in the same sequence).
3. **Novelty heads contribute little to task performance** (Sections 3 and 4). Novelty-ranked removal is less damaging than random removal on average. Additional tasks and training-stage analyses support this limited task contribution.
4. **Repetition heads track abstract representations and contribute modestly to task performance** (Section 4). Repetition heads correlate with format-invariant concept representation (CVs) and co-occur with CV heads across layers. Removing them impairs performance more than random removal, yet leaves the CV representation intact.
5. **Novelty-like attention mirrors human gaze shifts** (Section 5). Across all tested models, novelty-like attention follows pattern-dependent changes in human gaze allocation, whereas repetition-like attention shows the opposite relationship.

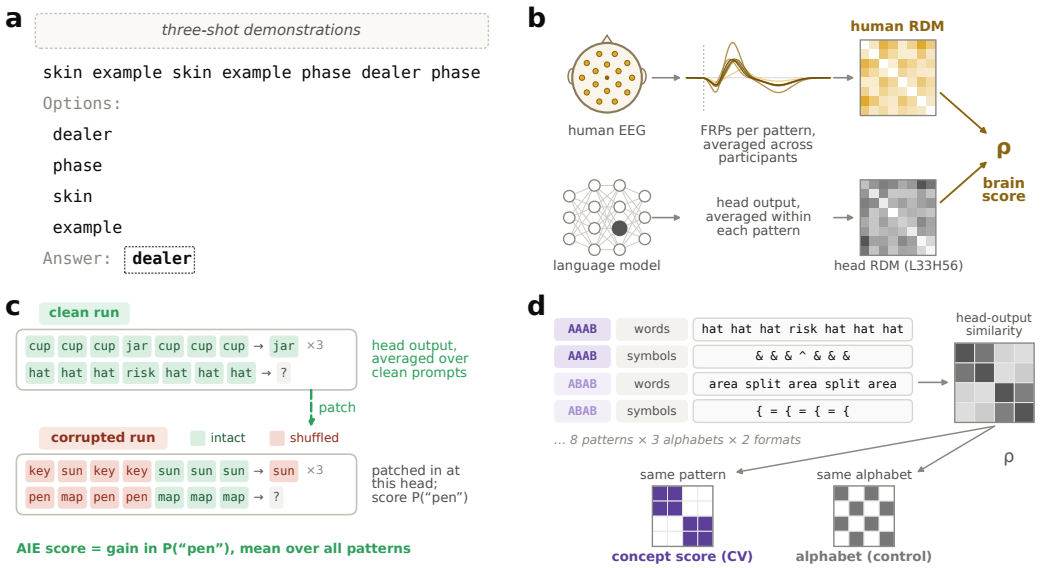

Figure 2: **The task prompt and the three head scores.** (a) A query as presented to the model, preceded by three solved demonstrations in the same format (not shown here). The seven query symbols are based on one of eight abstract patterns (here, `ABABCDC → D`) and the model's next token after "Answer:" is scored. (b) Brain score: Spearman correlation between a head's representational dissimilarity matrix (RDM), built from its output averaged within each pattern, and the human frontal-FRP RDM. (c) Patching score: a head's output, averaged over clean prompts, is patched into a corrupted prompt from a different item. In the corrupted prompt, the elements of the first block (e.g., `AAAB` in `AAABAAA`) of every in-context demonstration and of the query are re-drawn at random from the item's symbols (red). This destroys the pattern while the remaining positions stay intact (green), and the demonstration answers are random. The gain in the probability of the clean answer (here "pen", which completes `AAABAAAB`), averaged over items from all eight patterns, is the head's average indirect effect (AIE); the heads with the highest AIE are the function-vector (FV) heads. (d) Concept score: RSA of head outputs over 1,200 items that vary in pattern (8), alphabet (3; German, English, Symbols) and response format (2; open-ended, multiple choice), against a same-pattern design; a same-alphabet design serves as a control.

Current study. The abstract-sequence experiment of Pinier et al. (2025) combined EEG and eye tracking during 400 unique trials. Our brain-alignment target is an 8×8 representational dissimilarity matrix (RDM) of group-average frontal FRPs: rows and columns index the eight abstract patterns, and entries measure dissimilarity between their FRP waveforms. We additionally use human gaze data as a separate comparison of attention profiles. Human data recording, preprocessing, and RDM construction are detailed in Appendix A.1.

We first investigated Llama-3.1-70B base (Grattafiori et al., 2024) and then evaluated the generality of our findings in 16 additional models spanning 3B–72B parameters. Cross-model summaries include Llama-3.1-70B, giving 17 models in total. The cohort covers base and instruction-tuned Llama and Qwen variants, plus Phi-4 and DeepSeek-R1-Distill-Llama-70B. We present the initial and cross-model findings together, using the same human FRP target and noting protocol differences where relevant. Throughout the experiment, we use a text-based adaptation of the human task and raw prompts (i.e., no chat template) built with three solved examples.

The initial Llama-3.1-70B evaluation achieved 70.5% accuracy. Across the cohort, larger models generally performed better within Llama and instruction-tuned Qwen, but size alone did not determine success: Qwen2.5 base scored 74.5% at 32B versus 54.0% at 72B, while Phi-4 reached 82.0% at 14B. Instruction tuning improved Llama-3.1-70B performance but reduced it for Llama-3.1-8B under the same raw-prompt protocol. Thus, performance depends on model family and training variant as well as scale. These immediate-answer scores do not measure each model's best achievable performance or the benefit of extended reasoning.

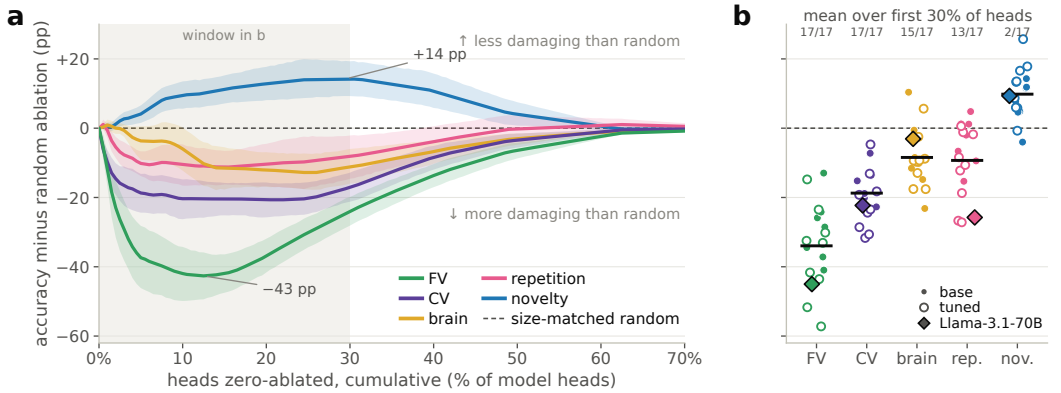

Figure 3: **Cumulative ablation relative to random removal, across 17 models.** Heads are zero-ablated cumulatively in each family's ranking order; values are accuracy minus the mean of five size-matched random ablations, in percentage points (pp), so negative values mean greater impairment than random. (a) Mean over models, with head counts interpolated on a percentage-of-heads axis; bands are model-bootstrap 95% intervals. FV-ranked ablations are the most damaging (-43 pp at 12.5% of heads), whereas removing novelty-ranked heads is less damaging than random removal (14 pp at 30%). (b) The same difference for each model, averaged over the first 30% of heads removed (shaded in a). Filled points are base models, open points instruction-tuned or distilled, and the diamond is Llama-3.1-70B; bars are means over models, and counts give the models below random controls.

2 Brain alignment versus causal importance

Does similarity to human brain activity identify the heads on which task performance depends, or those that represent the task's abstract patterns? We address this question first by comparing brain alignment with concept and patching scores across heads, then by testing the effects of removing heads selected by each criterion.

Three head scores. We scored every attention head by three criteria, illustrated in Figure 2b–d. The *brain score* (Figure 2b) measures how closely a head's pattern-level output geometry resembles the frontal-FRP RDM. The *patching score* (Figure 2c) estimates the average indirect effect on correct-answer prediction under activation patching, identifying candidate FV heads (Vig et al., 2020; Todd et al., 2024). The *concept score* (Figure 2d) measures correspondence with abstract pattern identity across presentation conditions (such as English vs. German words and multiple-choice vs. open-ended) using a procedure akin to that of Opiełka et al. (2026). In the cohort replications we estimated the activation-patching score by attribution patching for efficiency, as the two closely agreed in effect magnitude (Pearson $r = .983$) and selected 95% of the same top-20 heads in Llama-3.1-70B (Appendix A.3.4).

Brain scores do not predict patching scores. We paired each head's brain score with its patching score and computed Spearman rank correlations across all heads of a model. The comparison asks whether heads ranked higher for brain alignment also rank higher for causal effect. In Llama-3.1-70B, the two were only weakly related ($\rho = .03$), and the same held in every model (mean $\rho = .017$, range $[-.017, .077]$; Figure 9). Within a model, heads with higher brain scores therefore do not have larger patching effects.

Brain alignment relates more to concept representation, but varies across models. Brain–concept correlations averaged $\rho = .10$ across the 17 models (range $[-.20, .33]$), and were generally more positive and more variable than brain–patching correlations (Figure 8). The positive association observed in the exploratory Llama-3.1-70B analysis ($\rho = .23$) extended to every Llama model in the replication cohort ($\rho \in [.100, .325]$) and to DeepSeek ($\rho = .269$). Qwen showed a different pattern: associations were positive, though sometimes close to zero, in base models ($[.022, .206]$), but negative in all instruction-tuned variants ($[-.205, -.037]$). Phi-4 showed little association ($\rho = -.047$). Thus,

brain alignment often accompanies abstract-pattern representation, especially in Llama, whereas its relationship with patching effects remains consistently small.

Brain-ranked ablation is the least damaging. We zero-ablated heads cumulatively in each score's order, from highest to lowest, and compared the accuracy trajectory with matched random ablations (Appendix A.3.3; Appendix A.4.1). In Llama-3.1-70B, on the 200 English multiple-choice trials, patching-ranked removal dropped accuracy from 71.5% to below the 25% four-option chance level at 125 heads (2.44% of all heads) and to zero at 800 heads (15.63% of all heads). Brain-ranked removal stayed close to the random trajectory.

The same ordering held in all 17 models (Figure 3), but there brain-ranked removal was not equivalent to random: averaged over the first 30% of heads removed, it was more damaging than random in 15 of 17 models, against 17 of 17 for FV and CV. Peak accuracy loss beyond random ablation was 42.6 pp at 12.5% removal for FV-ranked heads, 20.7 pp at 19.5% for CV, and 12.8 pp at 24.5% for brain (peaks of the mean trajectories). Brain-aligned heads therefore carried some causal load, but far less than FV or CV heads, and the brain-ranked curve reached its peak only after removing far more heads.

3 Repetition- and novelty-like attention profiles

Do the heads most aligned with human FRPs share recognizable attention profiles, and do these profiles identify heads that matter for task performance? We first characterize their attention, then compare the resulting profiles and ablation effects across models.

Brain-aligned heads reveal repetition and novelty profiles. To characterize the highest brain-scoring heads, we looked at their attention on the seven elements making up each sequence, averaged over the 25 trials of each pattern type, giving an $8(patterns) \times 7(elements)$ profile per head. Clustering the top 20 brain-scoring heads by these profiles yielded three clusters of seven, nine, and four heads (Figure 4a; Appendix A.5.2). Two clusters showed distinctive attention patterns shared across their constituent heads: the novelty cluster (nine heads) preferentially attended to the *distinctive symbol*—the only symbol occurring once among the seven visible elements, a property present in five of the eight patterns—such as `B` in `AAABAAA` $\rightarrow$ `B`, whereas the repetition cluster (seven heads) emphasises repeated-symbol positions, such as the second `A` (Figure 4b; Appendix A.5.2). It should be noted that these names are used here to describe attention profiles, not established computational functions. The novelty label is supported by a direct attention measure: across all heads, similarity to the novelty template increased with the share of attention on the query's distinctive symbol ($r = .73$; Figure 4c). .

Template matching identifies carriers across models. We then averaged the profiles within each of these two clusters to obtain fixed templates and compared every head in the model with both templates. The comparison removes overall positional preferences (using a normalization and centering procedure further described in Appendix A.5.1) to emphasize how attention shifts across patterns. A head's *loading* is defined as its Pearson correlation with a template, while a *carrier* is a head whose loading exceeds the selection threshold of .50. Across heads, the two template loadings were strongly anticorrelated ($r = -0.83$), indicating contrasting attention organization rather than independent dimensions.

We then applied the frozen templates to every head of each cohort model. Averaged equally across models, repetition carriers accounted for 22.6% of the top 20 FRP heads versus 17.5% of all heads, indicating modest overrepresentation. Novelty carriers were common but not overrepresented (44.4% versus 46.7%, respectively), and 32.9% matched neither template (Appendix Figure 10).

Independent clustering recovers related profiles. To test whether similar profiles emerge without imposing the templates, we independently clustered the attention profiles of each model's top 20 FRP-aligned heads, yielding 3–13 clusters per model (methods in Appendix A.5.2). We then compared the three largest cluster means with the templates, finding a novelty-like profile ($r \geq .50$) in all 17 models and a repetition-like profile in 13. Repetition- and novelty-like attention therefore generalized across models under independent clustering, despite differences in the number and composition of clusters.

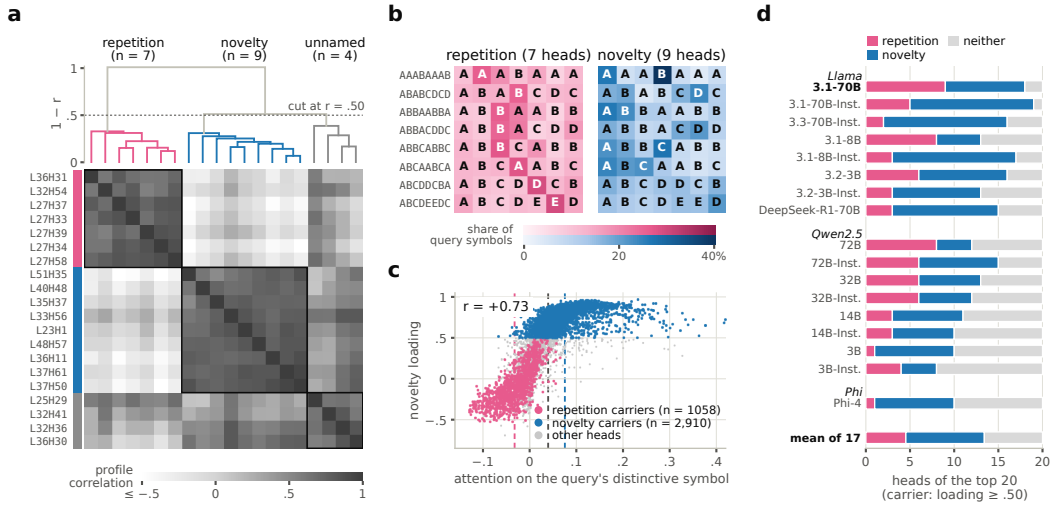

Figure 4: **Two attention profiles discovered among brain-aligned heads, and their recurrence across models.** (a) The 20 highest brain-scoring heads in Llama-3.1-70B, clustered by their attention profiles. Three clusters emerge: the two larger ones are named by their profiles, the third resembles both. (b) Mean attention profiles of the repetition and novelty cluster over the seven query-symbol positions for each pattern. These means are the templates used to identify carriers across models. (c) Label check on all 5,120 heads of Llama-3.1-70B: each head's novelty-template loading against its attention share on the query's distinctive symbol (the symbol shown once, averaged over the five patterns that have one). Dashed vertical lines mark the mean of the repetition carriers, of all heads, and of the novelty carriers. (d) For each of the 17 replication models, the number of its top-20 brain-scoring heads whose profile loads $\geq .50$ on the repetition template, on the novelty template, or on neither; the bottom row is the mean over models.

Repetition-ranked removal is more damaging than novelty-ranked removal. Removing repetition heads in descending order of template similarity produced greater overall accuracy loss than random-head removal, following a similar trajectory to the concept-ranked ablation, albeit not as pronounced. Surprisingly, removing novelty heads had much less effect: accuracy dropped to 63.5% after removal of 800 heads (15.63% of the model), compared with 71.5% before ablation, while the random-control median fell below 50%. The model therefore retained most of its task performance without this large set of novelty heads, whereas repetition-head removal was more disruptive.

The family ablations also generalized: in the cumulative sweeps (Figure 3), repetition-ranked removal was more damaging than random removal, with a peak excess loss of 11.2 percentage points at 15.5% of heads. Novelty-ranked removal, in contrast, never exceeded random removal by more than 0.3 points at any tested fraction, and over most of the curve was *less* damaging than removing the same number of random heads. The asymmetry between the two families observed in Llama-3.1-70B is therefore a property of the cohort, not of one model.

4 What the attention families do and do not do

The analyses in this section probed the two families beyond the primary-task ablations of Section 3 and were only run on Llama-3.1-70B and Pythia-1.4B. The latter was specifically used to study how these attention profiles and their effects on task performance evolved across training.

Repetition heads accompany the concept representation but do not supply it. In Llama-3.1-70B, two observations suggested that repetition heads might build the abstract-pattern representation carried by concept heads. First, across heads, repetition loading was positively related to concept score ($\rho = 0.45$; Appendix Figure 12a), whereas novelty loading was negatively related ($\rho = -0.19$). Second, the two sets co-occurred in depth: the 130 heads with the highest repetition loading lay in layers 16–35 (median 28) and the 150 highest-scoring concept heads in layers 22–39 (median

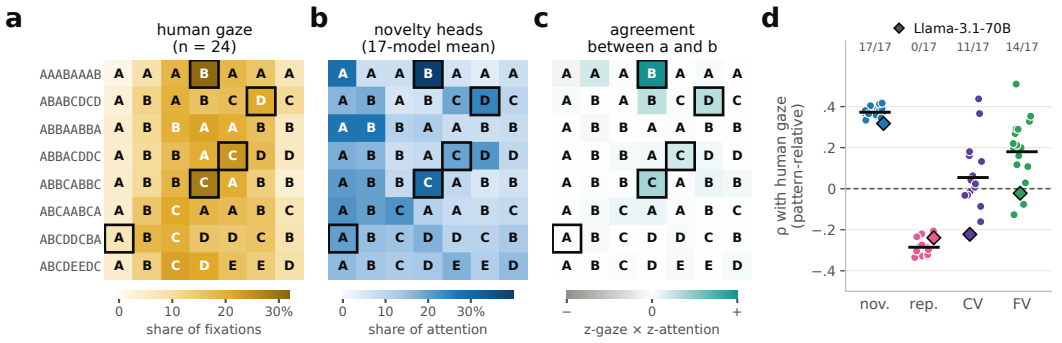

Figure 5: **Human gaze and head-family attention.** (a) Group fixation map of human participants over the seven visible query positions; each row is one pattern with its symbols overlaid, normalized to the share of that pattern's total fixation duration. (b) Attention of the top 1% novelty heads to the same positions, normalized per pattern and averaged over the 17 models. (c) Cell-wise agreement between (a) and (b): both maps are centered per position, z-scored, and multiplied, so the cells average to the Pearson correlation between (a) and (b). Boxes in (a–c) mark the five patterns whose only unrepeated symbol is the correct continuation; 70% of the agreement fell in these five cells and 95% in their rows. The first-position case (`ABCDDCB` → `A`) drew neither gaze nor attention. (d) Spearman correlation between the human map and the mean attention profile of each model's top 1% of heads per family, after centering both per position (pattern-relative). One point per model (diamond: Llama-3.1-70B); bars are means over the 17 models, and counts give the models with $\rho > 0$. Novelty heads correlated positively with gaze in all 17 models and repetition heads negatively in all 17.

33; Appendix Figure 12b). Did repetition-head removal therefore impair performance because it disrupted the representation of abstract patterns? It did not. Removing these 130 repetition heads lowered accuracy by 5.5 percentage points, but did not reduce cross-format concept RSA in later concept heads, and a classifier still identified the pattern from those heads' outputs with 100% accuracy (Appendix A.6.1; Figure 12c). Thus, the performance loss was not accompanied by a loss of the measured pattern information: repetition heads co-varied with the concept representation and occupied similar layers, but their removal did not eliminate the measured representation.

Novelty heads do not matter for other tasks, and did not earlier in training. The small effect of novelty-head removal might reflect specialization for another task. We therefore tested eight tasks spanning pattern completion, few-shot learning, prose, repeated-span verse, retrieval, coreference, and synthetic induction. Novelty-ranked removal was less damaging than the random median on seven tasks and fell within the random-control range on prose. In contrast, each task's own attribution ranking identified heads whose removal sharply impaired performance (Appendix Figure 14). The contrast therefore extended beyond the original pattern completion task: the tested tasks were sensitive to head removal, but novelty-template similarity did not identify the heads on which they most depended.

We also tracked a fixed set of novelty-like heads across 19 checkpoints of Pythia-1.4B (Biderman et al., 2023). Their attention increasingly resembled the final training profile. Measures of copying strength and attribution-based contribution to repeated-sequence prediction also increased, but remained below the reference thresholds used in that analysis. Separately, at the final checkpoint, removing the 32 heads most similar to the Llama novelty template reduced pattern-task accuracy from 42.5% to 39.5% (Appendix Figure 16). The growing prominence of this attention profile is therefore not matched by a growing role in the tested computations.

5 Relationship to Human Gaze

Novelty heads attend where humans look. Because the human FRPs are time-locked to fixations on sequence elements, we additionally compared head attention with human gaze. The summed duration of fixations on each sequence icon was summarized on the same pattern-by-position grid

as model attention (Figure 5a); Appendix A.7 documents the map construction. For each of the 17 models we took the top 1% of heads ranked by novelty loading, repetition loading, CV, FV, or FRP score, averaged their query-normalized 8×7 attention profiles, and correlated the result with the human map across the 56 cells. Novelty heads correlated positively with gaze in every model (mean $\rho = .373$, range $[.318, .421]$) and repetition heads negatively in every model (mean $\rho = -.285$, range $[-.335, -.206]$), while CV and FV heads fell between them on average, with mean $\rho_{\mathrm{CV}} = .054$ (range $[-.222, .438]$) and mean $\rho_{\mathrm{FV}} = .180$ (range $[-.127, .510]$), respectively (Figure 5d).

What the correspondence reflects. The task design confounds rarity with answer identity: in five of the eight patterns (`AAABAAAB`, `ABABCDCD`, `ABBACDDC`, `ABBCABBC`, `ABCDDCBA`) the only unrepeated symbol is also the correct continuation. The gaze correspondence was concentrated in these five (Figure 5c). Restricted to them, the pattern-relative correspondence of novelty heads averaged .47 across the 17 models, against .12 on the remaining three patterns. The repetition anticorrelation did not depend on this split ($-.34$ and $-.20$). Novelty heads and human gaze thus converged on the same cell, the single prior occurrence of the answer. Whether participants fixate it because it is the answer or because it is the odd one out cannot be decided from these patterns, since both accounts predict the same cell. What the ablations showed was that the model can attend to it without using it (Section 3). FV heads attended to the same symbol on the same five patterns (.28, versus .00 on the other three), so attention location alone does not separate idle heads from working ones.

6 Discussion

We tested whether similarity to human frontal FRPs identified attention heads that mattered for abstract-pattern completion. Across the tested models, brain alignment was a weaker guide to performance sensitivity than patching scores, yet revealed shared repetition- and novelty-like attention profiles.

Brain alignment is a weaker guide to performance-critical heads than causal scoring. Across 17 models, brain alignment was weakly associated with patching effects, and brain-ranked ablation was substantially less damaging on average than FV-ranked ablation. The mean excess loss beyond random peaked at 42.6 percentage points after removing 12.5% of heads for FV, compared with 12.8 points at 24.5% for brain-ranked removal. Brain-ranked removal therefore reached less than one-third of FV's peak excess loss despite removing almost twice the fraction of heads; CV-ranked removal had an intermediate effect (Figure 3). Brain alignment was not a consistent proxy for abstract-pattern representation either: its association with concept scores was positive in Llama and base Qwen models but negative in instruction-tuned Qwen models and Phi-4. The clearest common finding was therefore the weak link between brain resemblance and performance sensitivity, rather than a universal link between brain resemblance and abstraction .

Repetition and novelty profiles recur across models. Clustering brain-aligned heads separately in each model recovered attention profiles resembling the repetition and novelty templates, even though the templates were not used to form the clusters. The distinction between attending to repeated and distinctive sequence elements therefore extended beyond the original Llama analysis. These profiles did not exhaust the variety of brain-aligned attention: some clusters matched neither template. Their recurrence identified shared patterns of attention across models, but does not by itself imply that heads with similar profiles perform the same computations.

Repetition heads support performance while pattern information remains available. In Llama-3.1-70B, repetition loading correlated with concept scores, and the two head sets occupied overlapping layers. Yet removing 130 strongly repetition-like heads reduced accuracy while preserving the measured pattern information in later concept heads. These heads therefore contributed to performance without being essential for that measured representation under the tested ablation. Repetition-like heads might also account for some of the effects of brain-ranked removal, but the two rankings select different, partly overlapping sets. Ablating repetition-like, novelty-like, and other heads within the same brain-selected population would test this explanation directly.

Novelty-like attention: human resemblance with limited task effects. Novelty-ranked removal was less damaging than random removal on average, despite the profile's recurrence across models. The additional task battery and final-model Pythia ablation also showed limited performance sensitivity to novelty-ranked heads. Novelty-like attention also followed pattern-dependent changes in human gaze after removing overall positional preferences. This correspondence was concentrated in patterns where the distinctive symbol is also the correct answer, so sensitivity to rarity cannot be separated from attention to answer-relevant information. One possible explanation is that novelty-like attention is a *spandrel* (Gould & Lewontin, 1979)—a by-product of the model's architecture and training rather than a feature needed for these tasks. Softmax offers one reason this is possible: every head must distribute attention across the available tokens, even when its output has little influence on the answer. A recognizable attention profile can therefore coexist with a small task contribution. This does not explain why the profile favors distinctive elements, however, and our experiments do not establish its origin. Novelty-like heads resembled pattern-dependent changes in human gaze while contributing comparatively little to the tested tasks.

Relationship to other causal alignment studies. These findings need not conflict with evidence that training models to predict brain activity poorly impaired downstream performance (Merlin & Toneva, 2026), or that language-localized units were both causally important and more brain-aligned than random units (AlKhamissi et al., 2025). Those studies differ in selection criteria, representational targets, interventions, and behavioral endpoints. Changing representations through training or selecting language-responsive units does not test the same proposition as ranking attention heads by task-specific brain alignment.

Limitations and future directions. The main comparison used eight sequence patterns, fixed three-shot prompts, and one group-average frontal-FRP target. Here, "brain alignment" refers specifically to similarity with this target, not to correspondence with human brain activity in general. It captures one aspect of the EEG data, and other electrode groups, time windows, signal features, or recording modalities such as fMRI could identify different model components. Testing these alternatives, alongside additional tasks and prompting strategies, would establish how broadly the observed dissociation generalizes. Additionally, the human RDM gives participants unequal weight because they contributed unequal amounts of recording data and contains only 28 dependent pattern comparisons. Additional patterns, held-out participants, and alternative RDM constructions would strengthen the assessment of brain alignment. Regarding the LLMs, testing them more extensively on other tasks and prompting strategies would assess the scope of the findings. In particular, we did not identify a substantial task-specific role for novelty-like heads in the tested settings, but this does not establish that they lack a causal role elsewhere. Testing their contribution to other computations and tasks is an important direction for future work. Finally, small zero-ablation effects may reflect overlapping contributions from other heads, and layer-matched controls would help separate head selection from network location.

Conclusion. In this study, we combined representational comparisons with head ablations to examine the relationship between brain alignment and causal importance. Across the tested models, brain alignment was a substantially weaker guide to performance-critical heads than patching scores. Brain-aligned heads nevertheless revealed recurring repetition- and novelty-like attention profiles with different relationships to performance: repetition-ranked removal was more disruptive, whereas novelty-like attention more closely resembled pattern-dependent changes in human gaze. Together, these findings show that resemblance to human representations and attention does not reliably identify the model components most important for producing a correct answer.

Ethics statement

The EEG and eye-tracking data analyzed here are drawn from the de-identified dataset described by Pinier et al. (2025). This work reanalyzed this data and did not collect any additional human data. We release or share only de-identified data and derived representations.

Reproducibility statement

Code and data necessary to reproduce the results will be made publicly available upon publication.

AI USE STATEMENT

Generative AI tools were used for code assistance and for preliminary drafting and editing of prose. They were not used to generate human-participant data, to make analytic decisions without author review, or to fabricate experimental results, citations, or figures.

REFERENCES

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. 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. 10887–10911. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.544. URL https://aclanthology.org/2025.naacl-long.544/.

Richard Antonello and Alexander Huth. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79, 2024. doi: 10.1162/nol_a_00087.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Jeffrey S. Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E. Hummel, Rachel F. Heaton, Benjamin D. Evans, Jeffrey Mitchell, and Ryan Blything. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023. doi: 10.1017/S0140525X22002813.

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

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

Stephen Jay Gould and Richard C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B*, 205(1161):581–598, 1979. doi: 10.1098/rspb.1979.0086.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Olivia Guest and Andrea E. Martin. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6:213–227, 2023. doi: 10.1007/s42113-022-00166-x.

Gabriele Merlin and Mariya Toneva. When language models lose their mind: The consequences of brain misalignment. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/864b96035c34651e12202f8cd1d5e945-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86b3697c4eb7792c951831636bfdacd5-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2508.10057.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007. URL https://www.biorxiv.org/content/10.1101/407007v1.

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

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Daniel L. K. 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. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4060707/.

A SUPPORTING METHODS AND RESULTS

This section follows the main-text sequence: human data and model evaluation, head scoring and ablation, attention families, extended interventions, and gaze comparisons.

A.1 HUMAN STUDY AND FRONTAL-FRP TARGET

Participants and experimental setup. The source study (Pinier et al., 2025) recruited 25 adults with normal or corrected-to-normal vision and no reported personal or family history of epilepsy. Five withdrew before completing all five sessions, their available sessions were retained in the original study. One thing to note is that these recruitment counts are distinct from the participants with usable FRP RDMs in the present reanalysis. The planned experiment comprised 400 unique trials over five sessions (80 per session, and 50 trials per pattern overall). Unlike the model prompts, the human trials did not contain three solved demonstrations, but participants performed three trials with feedback at the begining of their first session.

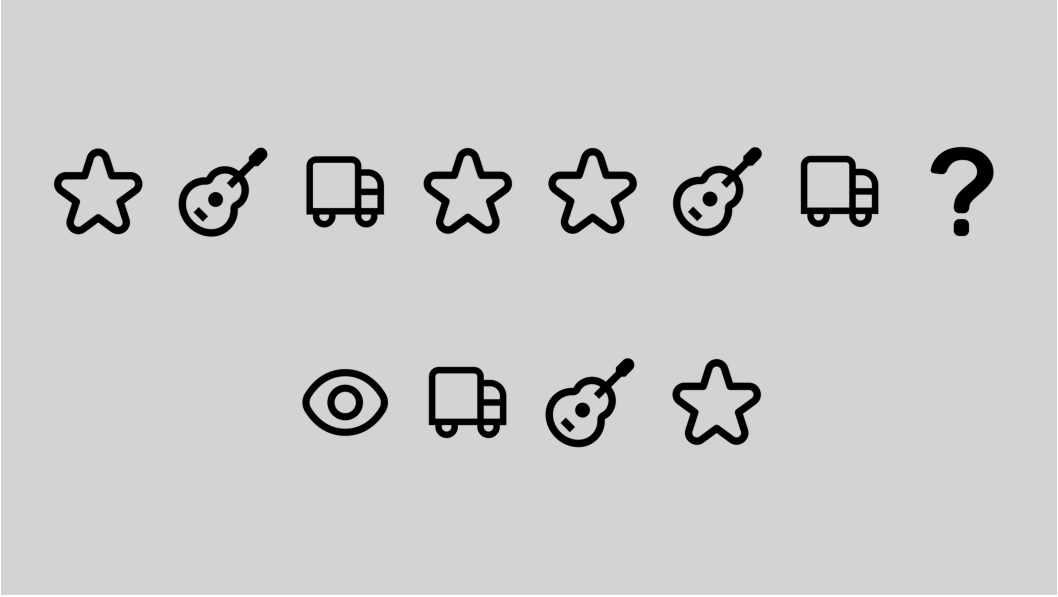

Figure 6: **Example trial from the human task.** Seven icons instantiate an abstract pattern; the participant selects the missing eighth icon from four alternatives. Here the pattern is `ABCAABCA`, and the star completes it. The FRP analysis retains fixations within the eight sequence-position regions, including the question-mark position, but not the response-option regions. Adapted from Pinier et al. (2025).

Recording and EEG preprocessing. The source study recorded 64-channel BioSemi EEG at 2,048 Hz and dominant-eye position with an EyeLink 1000 Plus at 2,000 Hz. Its documented MNE preprocessing marked bad channels, average-referenced the remaining channels, and applied zero-phase FIR notch filters at 50, 100, 150, 200, and 250 Hz. Extended Infomax ICA was fitted on a 1–100 Hz filtered copy; components not classified as brain by ICLabel were rejected, and the solution was applied to the notch-filtered data. A final 0.1–100 Hz bandpass, spherical-spline interpolation of bad channels, and average re-referencing completed preprocessing.

Fixation-related-potential construction. The frontal fixation-related-potential (FRP) signal was derived from the recordings reported by Pinier et al. (2025): for every retained fixation whose mean gaze position fell within one of the eight sequence-position bounding boxes (including the question mark), a 600-ms epoch of continuously preprocessed EEG beginning at fixation onset was retained, without fixation-specific baseline correction. Fixations on answer options were excluded. Within a trial, the retained epochs were averaged, so an FRP is a summary of self-paced inspection of the sequence rather than the response to one experimentally imposed event. Electrodes in the frontal

region of interest were then averaged at each time point to yield a trial-level temporal feature vector. This choice is substantively important: the target can reflect visual inspection, attention, and task difficulty as well as reasoning-related processing. Consecutive fixation epochs can overlap when their onsets are less than 600 ms apart.

Frontal electrode selection. The frontal region was an anatomical channel group, rather than electrodes selected for their correlation with model outputs. The analysis configuration specifies 17 channels (Figure 7): `Fp1, Fpz, Fp2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8`. These frontopolar, anterior-frontal, and frontal electrodes were averaged with equal weight at each time point. This reduces each pattern's response to a temporal waveform; the resulting RDM therefore compares waveform shapes, not spatial scalp patterns.

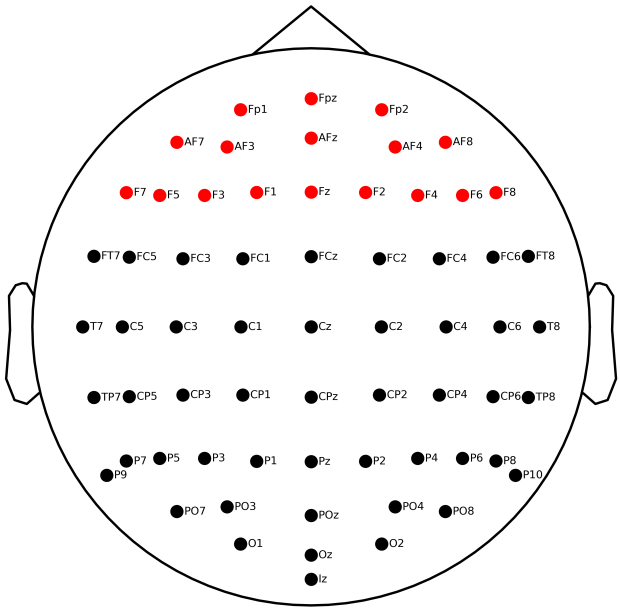

Figure 7: **Frontal electrodes used for the human FRP target.** The 17 electrodes listed in the text are highlighted in red; other displayed montage electrodes are black. The scalp is viewed from above, with the nose at the top. Adapted from Pinier et al. (2025).

The primary representational target. Our primary target is the original, group-average frontal-FRP RDM used in the initial Llama-3.1-70B discovery analysis. For each of the eight pattern classes, group-averaged FRP data were summarized as a pattern-level temporal feature vector; the dissimilarity of patterns p and q was their Pearson correlation distance, $D_{pq} = 1 - r(\mathbf{e}_p, \mathbf{e}_q)$. This produces one symmetric 8×8 RDM and hence 28 unique pattern-pair values. Thus, the target retains the contribution of every eligible recording session in the grand average before computing the RDM; participants with more retained data can contribute more to this estimator.

A.2 Model prompts and evaluation

Prompt formats and evaluation conditions. The specification below describes the cross-model replication prompts. Table 1 lists the surface alphabets, response formats, and trial counts used for each analysis stage.

Abstract patterns and item generation. The model task contains eight length-eight abstract patterns, `AAABAAAB`, `ABABCDCD`, `ABBAABBA`, `ABBACDDC`, `ABBCABBC`, `ABCAABCA`, `ABCDDCBA`, and `ABCDEEDC`. The eighth element is withheld from the query. For every item, abstract letters are mapped without replacement to surface tokens, yielding seven displayed sequence elements and

Stage	Model trials	Used to construct
Head RSA	1,200 trials across six conditions	CV and control scores
Brain RSA	200 English MC trials (RSA subset)	human-RDM scores
Attention scan	200 English MC trials	fingerprints and templates
Attribution patching	200 English open-ended trials	FV/AIE ranking
Independent clustering	saved fingerprints	unlabeled attention groups
Set construction	saved scores	full rankings and intervention grid
Zero-ablation sweep	200 English MC trials	causal trajectories and metrics

Table 1: Stage-specific trial conditions in the replication pipeline. MC denotes multiple choice. The six conditions cross English words, German words, and symbols with MC and open-ended completion, with 25 items per pattern and 200 trials per condition. Brain scores reuse the English MC subset of the head-RSA dataset; the attention scan and ablation sweep use a separately generated English MC set.

one correct continuation. The generator creates 25 independently sampled items for every pattern, alphabet, and response format, with random seed 42. The three alphabets are punctuation-like symbols, German words, and English words; the two response formats are open-ended completion and four-alternative multiple choice. Distractors are sampled first from other values used in the mapping and then, if needed, from unused values in the same alphabet. The correct answer and three distractors are shuffled deterministically.

Three-shot prompt. Every prompt contains three solved demonstrations followed by one unsolved query. The demonstrations instantiate the *same abstract pattern* as the query but use fresh, non-overlapping surface tokens. No natural-language instruction, system message, or additional introductory example is present. An open-ended demonstration is rendered as a space-separated seven-element sequence followed on the next line by `Answer:` x; a multiple-choice demonstration inserts an `Options:` block containing four shuffled surface-token alternatives before its answer. Demonstrations are separated by blank lines. The query has the same structure but ends at the literal boundary `Answer:`, with no trailing space and no supplied answer. The scored continuation is the correct eighth surface token. The current cross-model replication used this prompt with raw text (no chat template), even for instruction-tuned models. This holds prompt construction fixed; tokenizer-specific word pools can change literal lexical items. Chat-template and alternative-instruction experiments were separate prompt studies and are not combined with the results reported here.

Tokenization and continuation target. Word pools are filtered separately for each tokenizer: an eligible word must be represented by one token both in isolation and when preceded by a space. The preflight stage checks every content pool, option, and correct completion; it also verifies the exact prompt–continuation boundary by tokenizing the prompt alone and the prompt plus candidate continuation. A trial is accepted only when the latter adds exactly one token without changing the prefix token sequence.

A.2.1 Evaluation conditions and performance

Table 2 lists the included models and their clean-task performance under the common prompt protocol specified in Appendix A.2.

Formats and scoring. Performance is evaluated on the trial conditions specified in Table 1. Our primary measure is *exact next-token accuracy*: whether the highest-probability token across the full vocabulary is the contextually resolved correct answer token. This is not accuracy restricted to the four options, nor an evaluation of extended reasoning or generated explanations.

Scale, family, and instruction tuning. Performance generally improved with size within Llama and Qwen2.5-Instruct, but size alone did not determine success: Qwen2.5 base peaked at 32B (74.5%) rather than 72B (54.0%), and Phi-4 reached 82.0% with 14B parameters. Instruction tuning improved the larger Qwen models and Llama-3.1-70B, but not the 3B Qwen or 3B/8B Llama pairs under

Model family and size	Base	Instruct / post-trained
Llama-3.2-3B	51.0	39.5
Llama-3.1-8B	62.5	59.5
Llama-3.1-70B	70.5	77.5
Llama-3.3-70B	—	82.5
Qwen2.5-3B	29.5	25.0
Qwen2.5-14B	52.5	75.0
Qwen2.5-32B	74.5	77.5
Qwen2.5-72B	54.0	87.5
Phi-4 (14B)	—	82.0
DeepSeek-R1-Distill-Llama-70B	—	63.0

Table 2: **Clean English multiple-choice performance.** Exact next-token accuracy (%) on 200 trials per model under raw, three-shot prompting. Each populated cell is one of the 17 models. Dashes indicate no included model, not missing trials. Phi-4 and DeepSeek-R1-Distill-Llama are post-trained models; their placement in the right column does not imply identical training procedures.

this raw prompt. DeepSeek's 63.0% measures immediate completion, not the benefit of allowing a reasoning continuation. These are descriptive comparisons under a fixed prompting protocol, not isolated causal effects of parameter count or post-training.

Sensitivity to format. Six-format accuracy evaluations are available only for a subset and are kept separate from the 17-model sweep. In standalone raw-prompt evaluations of Llama-3.1-8B and 70B base, English open-ended accuracy was 91.5% and 90.0%, versus 62.5% and 68.5% for English multiple choice; German-word performance showed the same direction, while symbol conditions yielded only 18.5–33.0% exact next-token accuracy. These evaluations regenerate the full six-condition set, so their English multiple-choice items need not match the sweep items. Prompt formatting also mattered: in a paired Llama-3.3-70B-Instruct evaluation, adding the chat template raised English multiple-choice accuracy from 84.5% to 91.0% and symbol multiple-choice accuracy from 33.5% to 80.5%. Thus, low performance in the primary condition does not establish an absence of abstract-pattern competence. Subsequent causal comparisons use each model's own clean baseline under the shared English multiple-choice protocol.

A.3 Brain, concept, and function-vector scores

A.3.1 Head outputs, concept score, and control RSA

For each prompt, the pipeline extracts the input to the attention output projection at the final prompt position. This tensor is the concatenation of the individual attention-head output vectors before they are mixed by the layer's output projection; it is reshaped into one vector per layer and head. For each head, all trial vectors in the head-RSA dataset (Table 1) are ℓ_2-normalized and their pairwise cosine similarities are computed.

The *concept score*, used to rank candidate concept-vector (CV) heads, is Spearman's ρ between these head-output similarities and a binary design matrix whose entries indicate whether two items instantiate the same abstract pattern. The format and alphabet controls are computed identically, using same-format and same-alphabet design matrices. These are continuous rankings over every head, not intrinsically thresholded head families; a "top-k CV set" means the first k heads in the descending concept-score ordering. The use of all alphabets and both response formats broadens the criterion beyond one condition, but does not prove format invariance or causal use.

Head RDMs and brain scores. For each head, outputs from the English multiple-choice subset (Table 1) were ℓ_2-normalized trial by trial and averaged within each pattern. Writing the mean for pattern p as $\bar{\mathbf{v}}_p$, the between-pattern RDM entries were $D_{pq} = 1 - \bar{\mathbf{v}}_p^{\mathsf{T}} \bar{\mathbf{v}}_q$. The pattern means were not normalized again: their dot product equals the mean cosine similarity across all trial pairs from the two patterns, which is how the implementation computed these entries. A head's brain score was Spearman's ρ between the 28 upper-triangular entries of this RDM and the corresponding entries of the group-average frontal-FRP RDM (Appendix A.1).

A.3.2 FUNCTION-VECTOR SCORE BY ATTRIBUTION PATCHING

The function-vector (FV) ranking is estimated on 200 English-word, open-ended trials. A clean forward pass first averages each head's final-position output vector over the complete clean dataset. The corrupted dataset preserves vocabulary, prompt scaffold, response format, and target token, but independently randomizes the diagnostic prefix of every demonstration and query. Corruptions are rejected if they preserve the clean canonical pattern; demonstration answers are replaced by plausible values from the corrupted sequence. This destroys recoverable abstract-pattern information while avoiding a wholesale change of surface vocabulary.

On the corrupted prompts, the pipeline computes the gradient of the original correct answer token's probability with respect to each head's final-position output. The first-order attribution for a head and trial is the inner product of this gradient with the displacement from the corrupted activation to that head's clean mean activation. Averaging over trials gives the estimated average indirect effect (AIE). Heads are ranked in descending mean AIE to form the FV ordering. Model parameters are frozen; gradients are retained through activations by rooting autograd at the input embedding, and a loss scale is divided out before scores are saved. Runs containing non-finite probabilities or attributions are rejected. This score is an efficient first-order approximation to mean activation patching, not the cumulative zero-ablation effect used later.

A.3.3 CONCEPT AND FUNCTION VECTORS ON THE PATTERN TASK

The CV and FV rankings introduced in Section 2 define comparison sets at explicitly stated sizes, rather than categorical head types. The replication calculations are detailed in Appendices A.3.1 and A.3.2; the latter uses attribution rather than full activation patching. Their overlap is calculated by exact layer–head identity, and their causal consequences are compared on the common cumulative zero-ablation grid. This design separates three questions that are otherwise easy to conflate: whether a head's output geometry represents pattern identity, whether restoring that head is predicted to increase correct-answer probability, and whether removing a growing set of such heads actually changes behavior.

A.3.4 ATTRIBUTION PATCHING VALIDATION

In the exploratory analysis of Llama-3.1-70B, first-order attribution patching was compared with full single-head mean activation patching on the same trials and prompt configuration. Across heads, effect estimates had Pearson $r = .983$ and Spearman $\rho = .60$; the top-20, top-50, and top-150 head sets overlapped by 95%, 90%, and 91%, respectively. The strong agreement in effect magnitude and high-effect selections supports attribution patching as an efficient FV-selection method here. The lower overall rank correlation means that the complete rankings are not interchangeable, and this single-model validation does not establish equivalent accuracy on every replication model.

Attribution patching and cumulative zero ablation also test different interventions: the former approximates replacing one corrupted activation with its clean mean, whereas the latter removes sets of heads from clean prompts. We therefore use the sweep, rather than the patching score alone, to measure the behavioral consequences of the FV selection.

A.4 CUMULATIVE ABLATION AND CROSS-MODEL SCORE COMPARISONS

A.4.1 RANKINGS AND CUMULATIVE ZERO ABLATION

The set-construction stage creates five complete descending head orderings: FV (mean AIE, or full activation patching when explicitly available), brain (the selected primary human RDM), CV (concept RSA), repetition loading, and novelty loading. Non-finite values are placed last. Five seed-0 uniform random permutations of all heads form the standard descriptive reference band. The cumulative grid is $1, 2, 3, 5, 8, 13, 20, 32, 50, 80, 125, 200, 320, 500, 800, 1250, 2000, 3200, 5120$, truncated to values below the model's total head count and followed by that total. Thus each point removes a progressively larger set from the top of one fixed ranking; the random reference is five nested random trajectories, not a fresh sample at every head count.

Clean and ablated performance are evaluated on the same sweep trial set (Table 1). For every ordering and head count, the selected heads' slices in the input to each attention output projection are set to

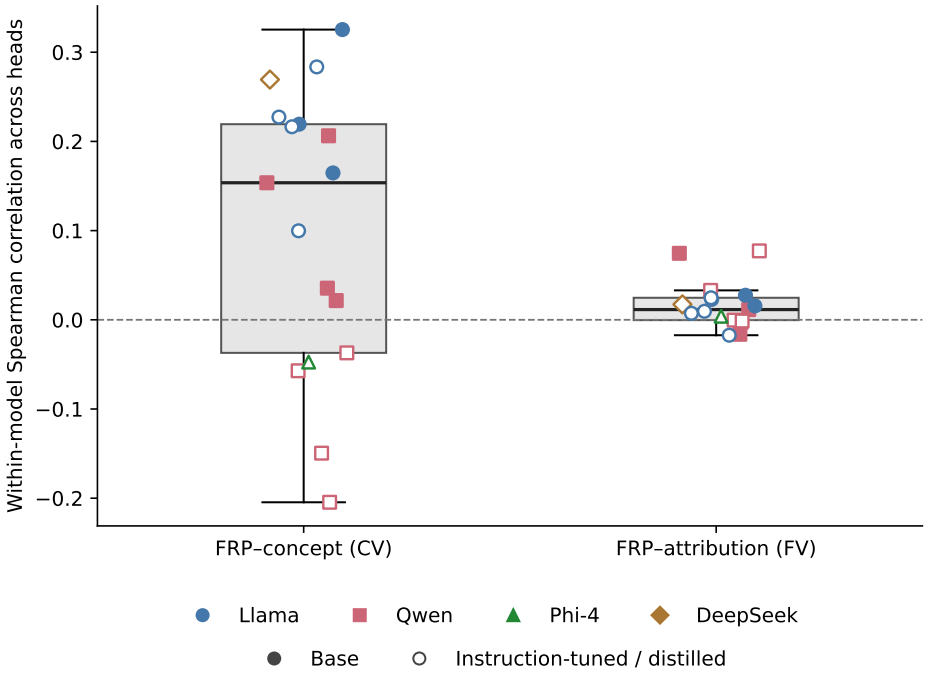

Figure 8: **Brain alignment and head-selection scores.** Each point represents one model's Spearman correlation across heads between frontal-FRP RSA and concept-RSA (CV) or attribution-patching (FV) scores; colors identify model families. Boxes show medians and interquartile ranges, with whiskers extending to the most extreme observations within 1.5 interquartile ranges. All model points are displayed. Undefined head scores are excluded pairwise. The 17 models are not independent samples of model families; these descriptive correlations are not participant-level inference.

zero at *every token position*. Slices are computed from the actual output-projection input width and number of attention heads, which also accommodates grouped-query architectures. The model is otherwise unchanged. All rankings necessarily converge when every head is removed, providing an implementation check. The primary behavioral outcome is exact next-token accuracy over the full vocabulary: the model's largest final-position logit must be the contextually resolved correct continuation token. The sweep output additionally stores the correct-token logit and probability, log-normalizer, top-two logits, full-vocabulary argmax, KL divergence from the clean next-token distribution, all four option logits, option-restricted prediction and accuracy, item pattern, and clean baselines. These arrays support per-pattern and alternative-metric analyses without rerunning the model.

The five standard random trajectories matched the ablated head count but not the layer distribution.

A.4.2 Cross-model summaries

Head-score associations were Spearman correlations across heads within each model, excluding undefined scores pairwise (Figures 8 and 9).

For the cumulative sweeps, head counts were expressed as percentages of each model's total heads and trajectories were linearly interpolated between tested counts. We subtracted each model's mean random-control accuracy and averaged the resulting differences equally across models. Negative values in Figure 3 therefore indicate greater impairment than random removal. Reported peak excess losses are maxima of the equal-model mean loss curves, not averages of model-specific maxima. Model-bootstrap intervals summarize variation within this cohort; related models are not independent samples of model families.

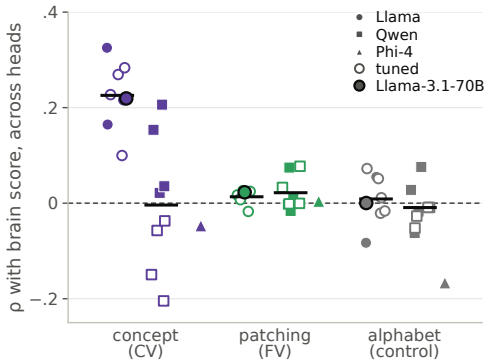

Figure 9: **Brain score against concept, patching and alphabet scores.** Per cohort model, the Spearman correlation across heads between brain score and concept (CV), attribution-patching (FV) or alphabet (control) score. Circles: Llama, including the DeepSeek distill; squares: Qwen2.5; triangle: Phi-4. Open markers are tuned models, the outlined circle is Llama-3.1-70B, and bars are family means. The brain–concept correlation was positive in every Llama and base Qwen model and negative in every tuned Qwen model and Phi-4; the other correlations stayed near zero.

A.5 Attention profiles, templates, and brain-selected clusters

A.5.1 All-head attention scan and frozen family loadings

The attention scan saves every head's attention row at the final prompt position for the trials specified in Table 1. Token spans are reconstructed from the rendered prompt, and the seven positions occupied by the unsolved query sequence are selected. Attention rows are first averaged over the 25 trials of each pattern. Each head's seven query-position values in this pattern mean are then divided by their sum, producing an 8×7 matrix. In the shared fingerprint implementation, averaging precedes query normalization. Finally, each position's mean across the eight patterns is removed. Flattening this matrix gives a 56-dimensional, pattern-centered attention fingerprint. Heads that place effectively no attention on the seven-symbol span have undefined template loadings and are sorted last in template rankings.

The frozen templates are the two discovery-cluster means described in Section 3; they are not refitted in the replication. For every head in every model, a family loading is the Pearson correlation between its pattern-centered fingerprint and the corresponding frozen template. Cross-model descriptive analyses use a common carrier cutoff of .50. Cumulative template ablations do not stop at the carrier cutoff: they rank *all* heads by loading and remove increasingly long prefixes.

A.5.2 Independent attention-profile clustering

For the top-20 analyses, pairwise distances between flattened attention profiles (Appendix A.5.1) were defined as $d = 1 - r$, where r is Pearson correlation. Average-linkage hierarchical clustering merged the groups with the smallest mean pairwise distance; the dendrogram was cut at distance .50. This cut does not require every pair of heads within a cluster to have $r \geq .50$. The three largest cluster means were compared with the frozen templates after clustering, using $r \geq .50$ to identify a match.

A.5.3 Template carriers, independent clusters, and membership agreement

Carrier prevalence. For each model, we computed carrier proportions among all heads and among its top 20 FRP heads using the cutoff in Appendix A.5.1. Overrepresentation was the within-model difference between these proportions; cohort summaries weighted models equally (Figure 10).

Membership agreement. Each template-carrier set was restricted to the same top-20 candidate pool and matched to the independent cluster with maximum Jaccard overlap (intersection divided by union). Mean overlap was .745 for repetition and .742 for novelty; exact set equality occurred in 4 and 4 of 17 models, respectively. These post-hoc best matches measure set agreement, not independent predictive validation.

Figure 11 shows the cluster profiles and their post-hoc template correspondence.

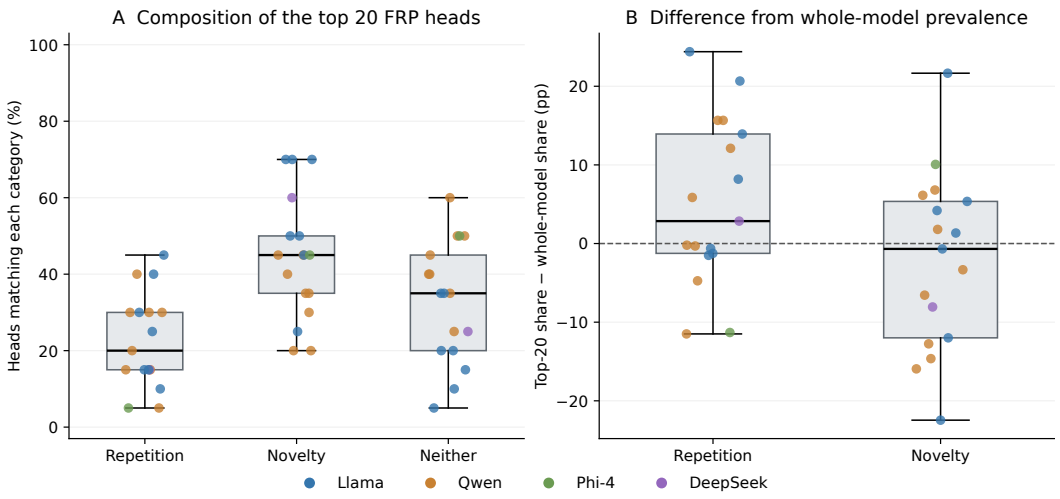

Figure 10: **Template carriers among the top 20 FRP heads.** Each point is one of the 17 models; colors identify model families. Boxes show medians and interquartile ranges, with whiskers extending to the most extreme observations within 1.5 interquartile ranges. (A) Percentage of the top 20 heads matching repetition, novelty, or neither template at $r \geq .50$. (B) Within-model difference between top-20 carrier share and its prevalence among all model heads; zero indicates equal prevalence. Positive values indicate overrepresentation, not a significance test.

A.6 Extended tests of attention-family function

These extensions examine representation, task performance, training, and pruning beyond the standardized replication sweep.

A.6.1 Repetition-head ablation and the concept representation

This analysis tested whether the repetition heads build the concept representation, given that the two were correlated across heads and lay in overlapping layers (Figure 12a,b).

Head sets. The *repetition set* contains the 130 heads with the highest repetition-template loading (all loadings $\geq .85$), spanning layers 16–35 (median 28). The *concept set* contains the 150 heads with the highest concept score, spanning layers 22–39 (median 33); nine heads belonged to both sets. All layers are 1-indexed.

Controls. Two controls accompany the repetition set. The *novelty control* contains 130 novelty carriers matched to the repetition set's layer profile; 103 heads required a nearest-layer substitute (median shift 3 layers, maximum 5). The *damage-matched random* sets are drawn from heads that carry neither template, with the repetition set's layer proportions. Their size was chosen by bisection so that their mean loss in correct-answer log-probability matches the repetition set's loss. This yielded sets of 25 heads, of which we drew five. The match was imperfect at this small effect size: the random sets cost more log-probability than the repetition set (0.090 vs. 0.059 nats) but less accuracy (3.8 vs. 5.5 points).

Readouts. Each set is zero-ablated in the English multiple-choice and open-ended formats (200 trials each), and we record the outputs of the concept heads at the final position. The readout heads are the concept-set heads in layers 34 and above that are not themselves ablated (73 heads; 72 for two random draws), so the measurement lies downstream of most of the repetition set. *Cross-format concept RSA* is computed per head as the Spearman correlation between its multiple-choice and open-ended item RDMs ($1-$cosine similarity), averaged over readout heads. It is compared with the unablated value on the same heads. The *pattern classifier* is a multinomial logistic regression over the concatenated readout-head outputs, trained on unablated multiple-choice trials (5-fold cross-validated accuracy 100%; permutation null 12.6%). It is then frozen and applied to the ablated outputs.

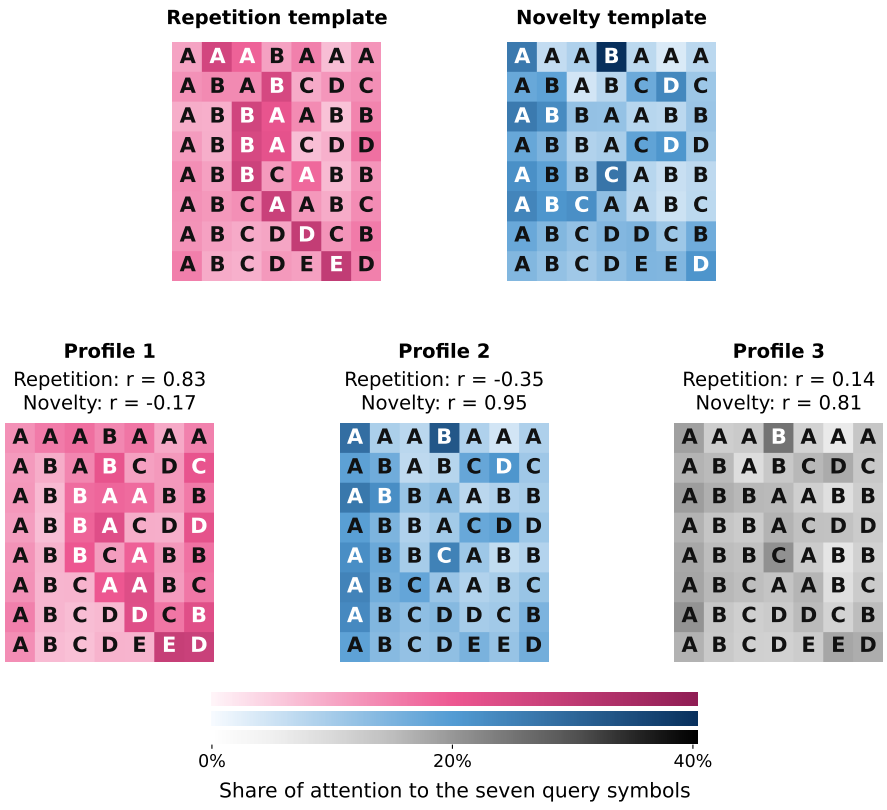

Figure 11: **Templates and independently discovered cluster profiles.** Top row: the original repetition and novelty templates. Bottom row: equal-model mean profiles of the three largest clusters among each model's top 20 FRP heads. Rows are patterns; columns are query positions, with the sequence symbols overlaid. Color intensity shows within-query attention share on a common scale; pink and blue link the first two aligned profiles to their display references, not to a categorical template assignment. For display, profiles were assigned one-to-one to the original cluster profiles by maximizing total Pearson similarity, without changing memberships. Title correlations compare each displayed mean profile with the indicated template using Pearson correlation after centering each position across patterns; they are not averages of model-wise correlations. This alignment is not independent evidence of recurrence or of three clusters per model. Post-hoc template comparisons found novelty-like matches in 17/17 models and repetition-like matches in 13/17 at $r \geq .50$.

Results. Removing the repetition set lowered accuracy from 70.5% to 65.0%, while concept RSA rose slightly from 0.401 to 0.427 (Figure 12c). The damage-matched random sets lowered accuracy by a mean of 3.8 points (range $[0, 8.5]$) and reduced concept RSA by 0.017 on average (SD 0.016). The novelty control lowered accuracy by 15 points and reduced concept RSA by 0.060. The classifier identified the pattern on 100% of trials in every condition. Removing the repetition heads therefore cost task accuracy without degrading the measured concept representation, and degraded that representation less than removing random or novelty heads.

A.6.2 Battery details

The earlier description identifies pattern completion in two formats (T1–T2), few-shot in-context learning (T3), prose (T4), repeated-span verse (T5), needle retrieval (T6), coreference cloze (T7), and synthetic induction (T8).

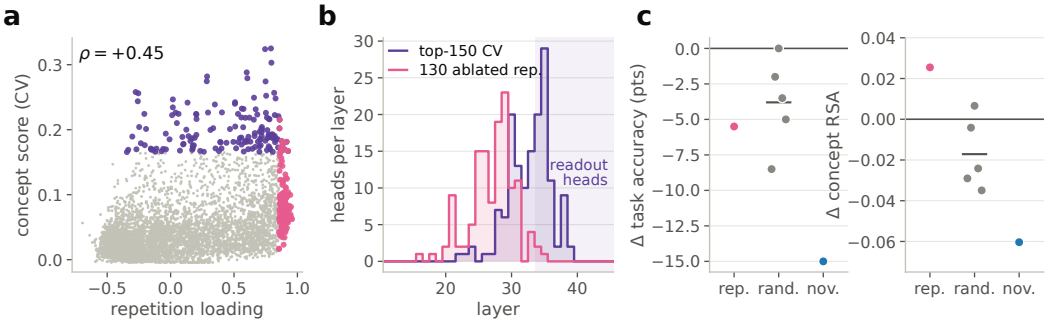

Figure 12: **Repetition-head removal reduced accuracy but preserved measured pattern information** (Llama-3.1-70B). (a) Repetition-template loading against concept score for all heads (Spearman ρ). Pink: the 130 ablated repetition heads; purple: the 150 highest-scoring concept (CV) heads; gray: all other heads. (b) Layer distribution of the same two sets. The shaded band marks the readout heads, the concept heads in layers 34 and above. (c) Change from the unablated model after zero-ablating the 130 repetition heads (rep.), five damage-matched random 25-head sets (rand.; bar: mean), or 130 layer-matched novelty carriers (nov.). Left: multiple-choice task accuracy. Right: cross-format concept RSA over the readout heads. A pattern classifier on the readout heads stayed at 100% accuracy in every condition.

A.6.3 Pythia trajectories

A.7 Human gaze and model attention

Gaze-map construction. The human gaze heatmaps were constructed by assigning fixations to sequence-item positions and summing the durations of all fixations on each icon. Trials are grouped by abstract pattern within each participant, pooling sessions before averaging; participant maps are then averaged with equal participant weight. The seven visible query positions are retained for comparison with the model profiles, excluding the eighth sequence position and answer options. Unfixated positions receive zero fixation duration within otherwise observed trials; trials without usable sequence fixations are marked missing. The original export includes eight sequence positions, whereas the subsequent comparison uses seven-position files.

A.7.1 Exploratory gaze comparison

The exploratory gaze comparison yielded template correlations .55/−.39 and participant means .41/−.28 for novelty/repetition. These are distinct from the top-1% replication comparisons shown in Figure 5. The human map construction is documented in Appendix A.7.

Gaze geometry and positional attention are different comparisons. Using the human gaze maps described in Appendix A.7, we performed two distinct comparisons. Comparing head-family attention RDMs with the human group gaze RDM over the seven visible positions yielded mean model-wise RSA of .494 for novelty carriers and .308 for repetition carriers. This tests which *patterns* have similar attentional organization, not whether models and humans attend to identical positions. Direct positional comparison instead used each model's top 1% of heads ranked separately by novelty, repetition, CV, FV, or FRP score—not the thresholded carrier sets above. Their query-normalized attention was averaged into an 8×7 profile and correlated with the corresponding human group gaze profile across 56 cells.

Raw positional correspondence was weakly negative for novelty and repetition (mean model-wise $\rho = -.051$ and $-.094$). After separately subtracting each position's across-pattern mean from the model and human profiles, novelty correspondence became positive (.373) and repetition correspondence negative (−.285), with these respective signs in all 17 models. CV, FV, and brain-selected groups showed smaller positive pattern-relative means (.054, .180, and .180; Figure 5d). Restricting selection to the top 0.5% retained the novelty/repetition pattern (.366/−.287). Thus, novelty-like heads resembled *pattern-dependent deviations* in gaze allocation despite differing in overall positional

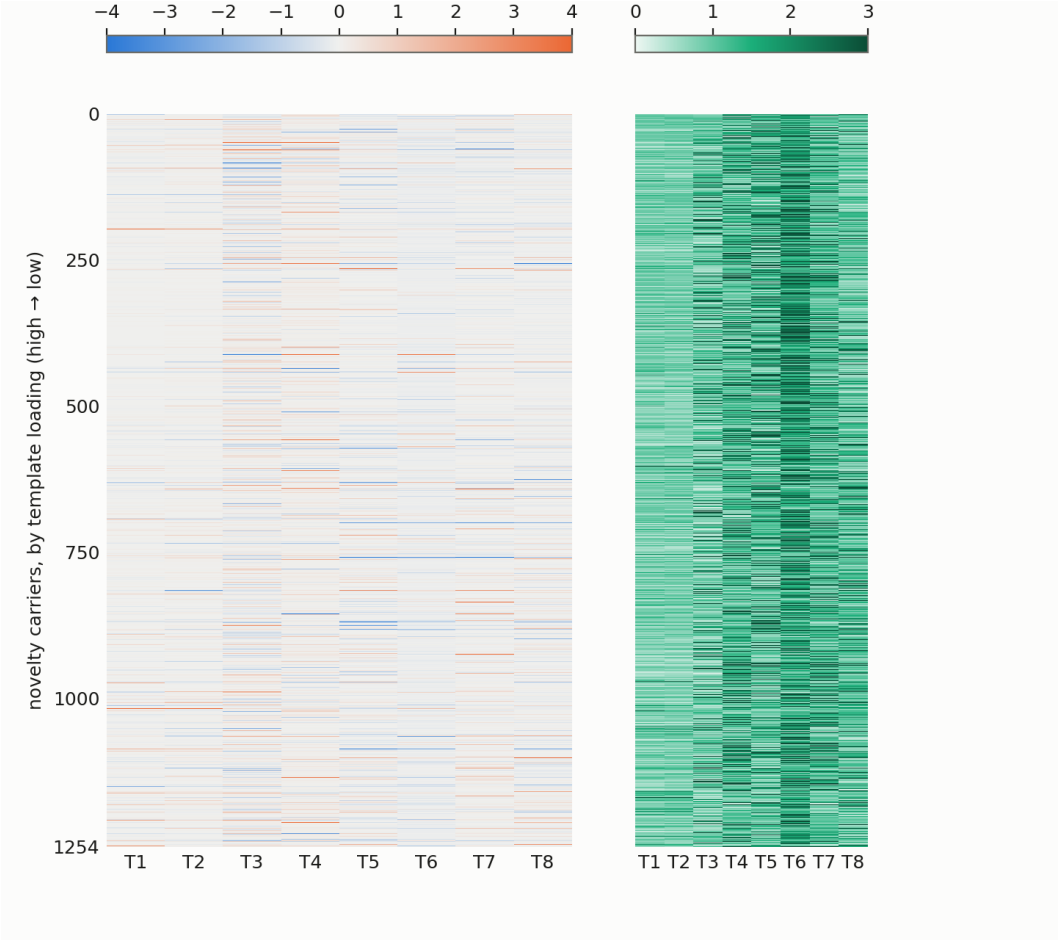

Figure 13: Attribution price of every novelty carrier on every task (z against layer-matched non-carriers) beside the gate column.

preference; repetition-like heads showed an opposing relative relationship. These normalized analyses do not measure how much total attention a model assigns to the sequence rather than prompt scaffolding, and mean correlations across models must not be confused with the correlation of a grand-average profile.

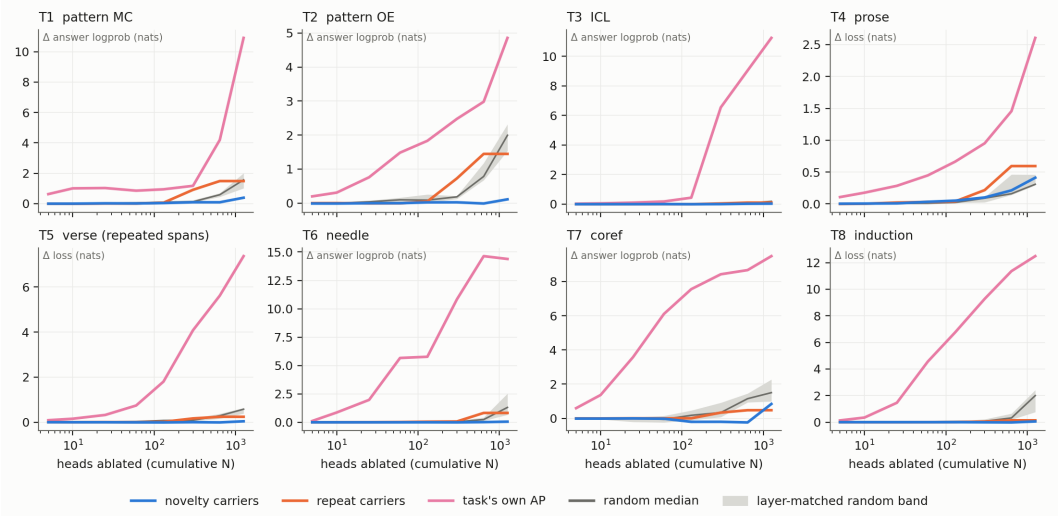

Figure 14: Exploratory eight-task cumulative-ablation trajectories.

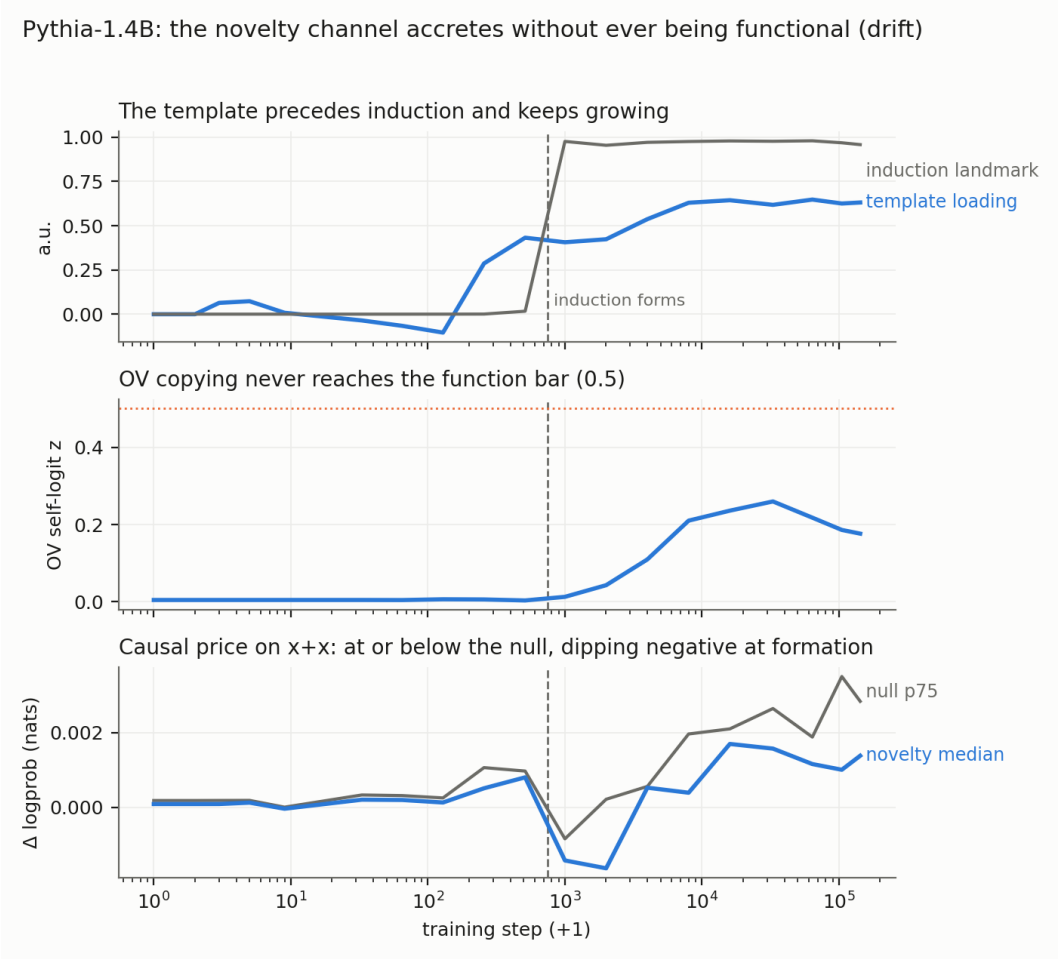

Figure 15: Template loading, OV copying and causal price of the future novelty population across Pythia-1.4B training.

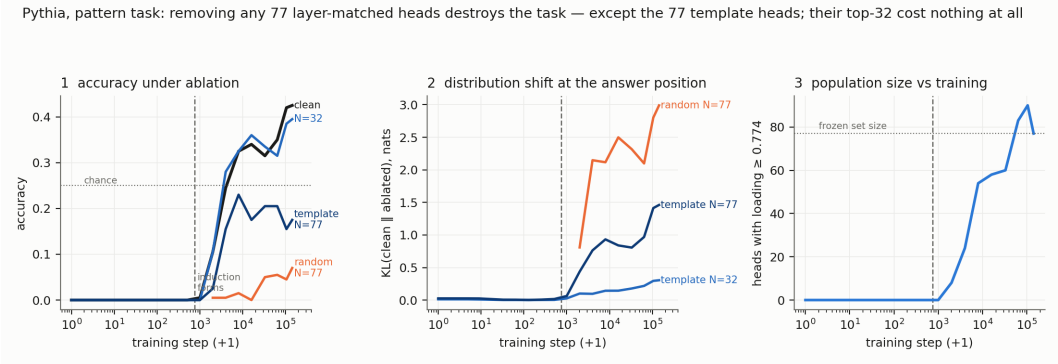

Figure 16: Exploratory Pythia pattern-task ablation, KL and template-population trajectories. These concern the tested interventions, not universal functionality.