nature portfolio

Peer Review File

Beyond representational alignment with brain-guided language models for robust reasoning

Corresponding Author: Professor Zhouchen Lin

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript presents a novel study exploring the causal utility of human neural representations in enhancing Large Language Models (LLMs). Moving beyond the traditional paradigm of correlational "representational alignment," the authors propose a "brain-guided" framework comprising Neural Activation guided Representation Intervention (NARI) and Fine-tuning (NARF). The study demonstrates that steering LLM representations toward directions derived from human fMRI signals during deductive reasoning tasks can correct model errors and improve generalization on out-of-distribution (OOD) tasks, such as propositional logic.

The work is innovative and timely, addressing the critical intersection of neuroscience and AI. The transition from observation (alignment) to intervention (guidance) represents a significant conceptual leap. However, to support the substantial claims regarding the superiority and robustness of "neural guidance," several critical issues regarding baselines, catastrophic forgetting, data scale, and the causal mechanism of the intervention should be addressed before the publication of the paper.

2. Major Concerns

（1）General Capabilities and Risk of Catastrophic Forgetting
I note a critical lack of evidence regarding the impact of the proposed methods (NARI and NARF) on the model's broader capabilities. The study relies on a limited amount of task-specific fMRI data, raising serious concerns about overfitting.

Catastrophic Forgetting: It is a common phenomenon in NeuroAI that aligning with biological signals or specialized fine-tuning can degrade performance on general tasks. It is unclear whether the observed improvements in deductive reasoning come at the cost of the model’s pre-existing language understanding, generation, or commonsense reasoning abilities.
Required Benchmarks: The authors must evaluate the NARF-tuned models on standard, broad-coverage benchmarks (e.g., MMLU, GSM8K, or general chat capabilities) to demonstrate that the model has not suffered from catastrophic forgetting. The effectiveness of the "regularization term" (Eq. 2) in preserving non-reasoning abilities must be empirically verified. Without this, it is difficult to rule out task-specific overfitting.

（2）Causal Interpretation and Intervention Logic
The methodological design relies on implicit assumptions that require empirical validation to establish a true causal link between neural alignment and reasoning performance.

Intervention on Correct Samples: The intervention is currently applied exclusively to cases where the model produces incorrect answers. This implies an assumption that correctly answered instances already possess "human-like" representations. This is neither justified nor validated. The authors must analyze the effects of neural-guided interventions on correctly answered samples. Does the intervention leave performance unchanged, or does it degrade correct reasoning trajectories? If correct responses are robust to intervention, the authors' claims are strengthened; if not, the causal link is weakened.
Mechanism of Signal Effectiveness: The reported noise ceilings of the reasoning-related brain regions are relatively low. A more explicit discussion is needed to explain how neural activations with limited signal-to-noise ratios can act as effective gradient directions or regularization signals.

Intercept Ablation: The finding that removing the intercept term b in ridge regression improves performance is counter-intuitive. Does this imply that the pattern (direction) of neural activity is informative, while the magnitude (baseline activation) introduces noise? A deeper discussion is warranted to clarify the mechanism of alignment.

（3）Data Scale, Robustness, and Cross-Subject Generalization
The study utilizes a very small fMRI dataset (<70 unique problems from 10 subjects) to train models evaluated on >50,000 synthetic test problems. This mismatch raises concerns about the universality of the extracted signals.

Leave-One-Subject-Out (LOSO) Analysis: The current method aggregates signals from all 10 subjects. To prove that the method captures a universal reasoning mechanism rather than overfitting to specific subjects, the authors must perform a LOSO analysis (train on N-1 subjects, evaluate on the held-out subject).
External Dataset Validation: Relying on a single dataset (Lytle et al., 2020) limits the robustness of the conclusions. Validation on an independent dataset (e.g., HCP Relational Processing Task or Pereira et al., 2018) is strongly recommended to demonstrate that the pipeline is not an artifact of one specific experimental design.

（4）Domain Shift and Failure Analysis
Pseudowords vs. Natural Language: The fMRI task uses pseudowords to isolate logic, but LLMs are pre-trained on natural language. Please discuss whether this distribution shift affects the efficiency of the guidance. Does the method work equally well on natural language reasoning datasets (e.g., FOLIO) compared to the synthetic pseudoword test set?
Granular Failure Analysis: While the authors analyze failure cases based on subject performance, a more granular analysis is needed. Are there specific linguistic characteristics (e.g., negation, sentence complexity) or logical structures where the fMRI signal consistently fails to steer the model?

3. Minor Concerns

（1）Computational Cost: Please clarify the additional computational latency introduced by the NARI (inference-time intervention) method. Is it feasible for real-time applications?
（2）Visualization: Figure 3 (Methodology) is somewhat cluttered. The distinction between the "Intervention" path and the "Fine-tuning" path could be made visually distinct (e.g., different color coding for data flows).
（3）Terminology: The terms "Deductive Reasoning Network" and "Logic Network" appear to be used interchangeably. Consistent terminology would improve readability.
（4）Statistical Significance: In the OOD generalization results (Fig. 6), please ensure that statistical significance tests are reported for the performance gaps between "Language SFT" and "NARF".
(5) Regarding Individual Differences and Insufficient Robustness Analysis: The paper notes significant variability in intervention success rates across subjects (Fig. S5), which correlates with subjects' behavioral accuracy. This suggests that the effectiveness of neural guidance may be highly dependent on the quality of individual neural representations. We recommend a further analysis to clarify whether the method is primarily effective only for signals from "high-performing subjects." Crucially, it should be investigated whether signals from subjects with lower alignment or performance could introduce noise and potentially degrade model performance. Addressing this is essential for assessing the generalizability and practical robustness of the proposed guidance approach.
(6) Regarding Temporal Resolution and Compatibility with Chain-of-Thought Models: The fMRI signals used are slow hemodynamic responses, while chain-of-thought reasoning is a dynamic, sequential process. Aligning these integrated fMRI signals with an averaged representation of the thinking tokens presents a potential neuro-temporal mismatch. We recommend explicitly acknowledging this limitation in the Discussion section. It would be valuable to discuss future directions that could address this, such as utilizing neuroimaging techniques with higher temporal resolution (e.g., MEG or EEG) to capture the fast dynamics of reasoning processes, potentially enabling more fine-grained alignment with intermediate reasoning steps.

(Remarks on code availability)
I did not find the source code in the supplementary.

Reviewer #2

(Remarks to the Author)
Summary

The paper explores whether large language models (LLMs) align with the neural mechanisms of human deductive reasoning and if such brain signals can be leveraged to enhance their reasoning performance. The authors compare internal model representations with task-fMRI data from humans solving logic problems and find that the best LLMs capture approximately 90% of the explainable variance in aggregate neural responses across deductive reasoning regions. They then build on this by showing that one can improve LLMs' reasoning capabilities by leveraging those human brain signals. Specifically, the authors introduced a brain-guided framework featuring two methods: Neural Activation guided Representation Intervention (NARI), which applies inference-time steering to the model’s representations, and Neural Activation guided Representation Fine-tuning (NARF), which internalizes neural knowledge into the model's parameters. Tested across 10 different LLMs ranging from 1.5B to 72B parameters, the authors find that their NARF approach complements language-only behavioral fine-tuning; applying both achieved absolute accuracy gains of up to 22% on logical reasoning tasks.

Strengths

S1. To our knowledge, this work provides the first empirical demonstration that task-evoked human neural signals can be leveraged to functionally enhance deductive reasoning in LLMs. This is a more causal confirmation of the to-date representational correlation observations.

S2. The reported ceiling values (approximately 0.25) are quite respectable for the reasoning domain of neural data, suggesting a somewhat robust signal-to-noise ratio in the brain-model mapping.

S3. The authors demonstrate the framework's broad applicability by validating it across a diverse suite of state-of-the-art architectures (including Qwen, Llama, Mistral, and Phi) and varying parameter scales.

S4. The experimental design for the NARI and NARF methods is well-designed, particularly in the selection of baseline controls to isolate the specific impact of neural guidance.

S5. The authors demonstrate that neural guidance provides benefits orthogonal to standard language-only supervision, specifically by refining "reasoning representations" in intermediate layers rather than just final output tokens.

Weaknesses

W1. While the authors frame this as a "paradigm shift" toward cognitively principled AI, the long-term scalability of the approach is questionable. Given the high cost and logistical difficulty of acquiring high-quality fMRI data, it is unclear if this method can truly compete with large-scale behavioral fine-tuning or synthetic data generation. The authors should clarify the "neural data-efficiency" of their model – specifically, how much brain data is required to achieve gains that cannot be replicated through standard machine learning techniques.

W2. The study's empirical breadth is limited by the use of a single fMRI dataset for both alignment and validation. While task-specific fMRI data for reasoning is scarce, the reliance on a singular source raises concerns regarding domain-generality. To establish NARI/NARF as robust NeuroAI tools, the authors should demonstrate their efficacy in other cognitive domains (e.g., language processing or speech), where larger, diverse neural datasets are more readily available. This would clarify whether the method captures a general principle of brain-AI integration or is specifically tuned to the unique structures of formal logic.

W3. The functional utility of NARI is partially obscured by the results in Figures 4h and 4k. For certain models, the performance gains over "Random Direction" baselines appear marginal. If a stochastic perturbation of the same magnitude performs comparably to a brain-derived direction, it suggests that NARI might be acting as a generic form of representation-space regularization rather than a specific cognitive guide. Further clarification is needed on how test accuracy scales as a function of the intervention parameter (α) for these random directions. The alignment of the brain-trained models actually also decreases in some cases (Fig. S8).

W4. To quantify neural predictivity, the authors take the maximum alignment across all evaluated layers as the final "brain-score," following Schrimpf et al. (2021). However, this approach has been iterated on since to resolve issues regarding the "oracle" nature of selecting the best layer on the test set itself which might inflate predictivity scores. Contemporary studies in NeuroAI mitigate this selection bias by either i) utilizing functional localizers [1] to isolate model units that are specifically selective for a given functionality (e.g., deductive reasoning), mirroring how functional regions are identified in the human brain; or ii) using a validation or "layer-selection" set to commit to a specific layer before evaluating performance on a held-out test set. These approaches ensure that the final evaluation is independent of the layer or unit selection process, thereby providing a more scientifically valid estimate of representational alignment. In our experience, with a sufficiently large dataset these results will be consistent but since this is a new domain of NeuroAI inquiry, we feel it is worthwhile to double-check.

W5. The remarkably high neural predictivity observed in untrained models (Figure 2e) is a potential "red flag" that may point to methodological artifacts. Our intuition is that randomly initialized weights should not, in principle, align closely with complex human reasoning. This might suggest train-test contamination within the k-fold cross-validation, potentially caused by shuffling samples with strong temporal or contextual dependencies. If samples from the same stimulus block appear in both sets, the regression may be fitting to low-level signal characteristics or scanner noise rather than genuine neural representations.

W6. The PCA and KDE visualizations (e.g., Figure 4c) primarily illustrate the consequences of fine-tuning (i.e., better class separation) rather than the mechanisms of neural guidance. These plots confirm that the model's accuracy has improved, but they do not sufficiently explain how the neural signals steered the model toward a more robust logical manifold.

Presentation inconsistencies

P1. There is a lack of uniformity in the models presented across the primary figures. Figure 4 focuses on Qwen2 and Llama2/3; Figure 5 adds Mistral; and Figure 6 introduces a wide array of Instruct-tuned models (Phi, Gemma, etc.). This shifting selection makes it difficult to assess the consistency and robustness of the results across the entire study.

P2. Figure S8 indicates that in some instances, the alignment of brain-trained models actually decreases. This counterintuitive finding deserves a more thorough discussion in the main text, as it complicates the narrative that alignment and performance are linearly coupled.

Questions

Q1. Was there a theoretical or empirical rationale for focusing the intervention exclusively on the outputs of the attention modules and layers, rather than including the outputs of the Feed-Forward Networks (FFNs) for example?

Q2. The results in Figure S5d are really interesting if significant. Could the authors provide a formal statistical test for the significance of this result and explain why such a notable finding was relegated to the Supplementary Information?

Signed,
Badr AlKhamissi & Martin Schrimpf

References
[1] AlKhamissi et al. (2025): From Language to Cognition: How LLMs Outgrow the Human Language Network. (EMNLP)

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the authors for the careful, well-organized revision and their detailed responses to each of my first-round concerns. The revised manuscript is substantially strengthened, in particular by (i) the broad-capability benchmarking on MMLU/GSM8K/HumanEval/MBPP, (ii) the external validation on the HCP Relational Processing task, (iii) the FOLIO results demonstrating transfer to natural-language reasoning, (iv) the subject-robustness analyses, and (v) the newly added paired statistical tests.
Below I comment on each of my previous points. Most are resolved; two would benefit from small, targeted additions before acceptance, and I also note a minor discussion that would enrich the paper.
Major:
1. General capabilities / catastrophic forgetting — Resolved.
The new evaluation of all fine-tuned models on MMLU, GSM8K, HumanEval, and MBPP (Table S1, Sec. S11.5) convincingly shows that NARF and NARF+Label do not degrade broader capabilities — fluctuations are small and several cells even improve. The clarification that the regularization term in Eq. (2) is a reasoning-specific, label-free device (not a general anti-forgetting mechanism) resolves my earlier ambiguity about its role.
At a more mechanistic level, I remain curious about why fine-tuning on neural guidance derived from a small set of problems leaves broader abilities essentially intact. A brief discussion of the factors the authors believe are responsible (e.g., LoRA restricted to attention modules in a few middle layers, small λ_narf, limited training samples, the directional nature of the objective) would enrich the paper. I do not consider this essential for acceptance.
2. Causal interpretation and intervention logic
2.1 Intervention on correct samples — Largely addressed; one small empirical check still recommended.
The conceptual argument in the new Fig. S13 / Sec. S12 — that human-like representations do not guarantee a "more correct" direction for already-correct cases, so intervention on such cases cannot serve as a clean causal test in the instance-specific setting — is reasonable, and I accept it. The population-level NARI(gen.) result on the full test set is reassuring in that applying the direction broadly does not destabilize model behaviour.
Nonetheless, I would still like a small empirical sanity check to complement the conceptual argument. Specifically, for the instance-specific intervention protocol, please run the intervention on a random sample of problems that the model already answers correctly and report how often it flips the answer to incorrect. A low flip rate would further strengthen the causal-link claim; a non-trivial rate would itself be valuable to

acknowledge. This is a modest Supplementary Information addition and does not change the main conclusions.

2.2 Signal effectiveness under moderate noise ceilings — Resolved.

The derivation showing that zero-mean noise cancels in $YcYc^\top$ under expectation (Sec. S5), combined with the functional-localization-based predictivity re-analysis, satisfactorily explains how useful guidance can be extracted from fMRI signals with modest per-voxel SNR.

2.3 Intercept ablation — Resolved.

The derivation $\nabla S_{w/o\ icpt} - \nabla S_{w/\ icpt} = W^\top b$ clearly shows that excluding the intercept adds a systematic-shift direction modulated by the joint model–brain structure, rather than introducing noise. The expanded discussion in Sec. S5 is clear.

3.2 LOSO — Accepted with a small caveat.

I accept the authors' argument that a literal LOSO is not directly applicable because the method does not predict individual subjects' neural responses at test time, and I agree that the subject-inclusion analyses in Fig. S11 (1→10 subjects, both high-to-low and lowto-high orders) adequately address the underlying concern about overfitting to specific subjects. I would simply recommend explicitly flagging in the text that Fig. S11 serves as their surrogate for LOSO and briefly explaining why the two speak to the same concern; this will help readers who are primed to look for a LOSO analysis and prevent the replacement from appearing to side-step the question.

3.3 External dataset validation — Fully resolved.

The new HCP Relational Processing experiments (Sec. S13 and Fig. S14), covering intervention, general-direction, and NARF+Label across five model families, constitute a meaningful external validation that the pipeline is not an artefact of a single experimental design. The drop in intervention success rate to ≈80% under the modality shift is properly acknowledged and reasonably interpreted as a boundary condition.

4. Domain shift and failure analysis

4.1 Pseudowords vs natural language / FOLIO — Resolved.

The FOLIO-wiki-curated-TF results in Fig. 6b demonstrate that guidance extracted from pseudoword stimuli generalizes to natural-language first-order-logic reasoning. The authors' acknowledgement that a natural-language-based guidance pipeline is currently blocked by the lack of suitable task-fMRI data is reasonable.

4.2 Granular failure analysis — Addressed, but quantitative support is still needed.

Figure S7 provides a useful stratification by affirmative/negation, 2-of-3 vs all-premises complexity, and syllogistic vs transitive type, and I accept the authors' observation that no consistent failure characteristic is apparent. However, the current conclusion of "no consistent failure mode" rests on visual inspection only. Please supplement Fig. S7 with per-subtype paired statistical tests comparing NARF(+Label) to its baseline across models and the five seeds, so that the negative result is supported quantitatively. This is a minor but important addition for the rigor of the claim.

Minor

5. Computational cost — Resolved. (Sec. S6: NARI(gen.) overhead negligible; perinstance NARI verification can be 1–40× forward-pass cost depending on difficulty.)

6. Figure 3 — Resolved. The Intervention and Fine-tuning paths are now visually separable.

7. Terminology — Resolved. "Deductive reasoning network" is used consistently.

8. Statistical significance in Fig. 6 — Resolved. The five-seed paired t-tests provide appropriate statistical backing.

9. Individual differences and signals from low-performing subjects — Fully resolved.

Figure S11d is particularly reassuring: including low-performing subjects does not degrade performance, which directly answers my earlier concern about whether noisier signals could hurt rather than help.

10. fMRI temporal resolution / MEG-EEG future direction — Resolved.

The limitation and the suggested direction (MEG/EEG for finer-grained, step-wise guidance) are now acknowledged in the Discussion.

11. Source code — Resolved. The GitHub repository is provided in the Code Availability statement.

I recommend acceptance conditional on two small additions that can be accommodated in the Supplementary Information:

(a) (Concern 2.1) An empirical sanity check: apply instance-specific NARI to a random sample of initially-correct problems and report the flip-to-incorrect rate.

(b) (Concern 4.2) Per-subtype paired statistical tests accompanying Fig. S7, so that the "no consistent failure mode" conclusion is quantitatively supported.

I would also welcome — but do not require — a brief mechanistic discussion (under Concern 1) of the design choices that plausibly underlie NARF's resistance to catastrophic forgetting, and an explicit note in the text that Fig. S11 plays the role of a LOSO equivalent subject-robustness analysis (Concern 3.2).

(Remarks on code availability)
The GitHub repository is provided in the Code Availability statement.I did not install and run the code.

Reviewer #2

(Remarks to the Author)
Thank you very much to the authors for their thorough response to the points that were raised.

W1. I agree that brain data can act as a complementary source of supervision, but my pushback was more on how much brain data we then actually need. I was hoping for something like scaling curves that have subject hours on the x-axis, and delta performance improvement over standard ML-training on the y-axis. Has the gain from brain data already peaked or do you predict it to continue scaling (log-linearly)? Sorry for not being clearer on this in the original review. I understand the author's concern over limited data availability however, so I am okay with this being left out.

W2. I am happy with the converging results on another dataset.

W3. Thank you for the new analyses. Great to see that random directions produce no change. It does seem like the alpha parameter is again quite specific to the model though, so this is important to highlight as a sensitivity of the method. The second part or your answer gives a key insight into how you think about the modeling, and I think this should be stated explicitly in the introduction: you view the neural data as a tool to improve core ML capabilities, but improved neural alignment is itself not the goal (it is for my research, so I incorrectly assumed it would be for you as well; hence I think good to clarify upfront).

W4. Using newer techniques to estimate neural predictivity slightly lowers scores, but maintains qualitative claims. I'm happy with this.

W5. Seems like the newer techniques from W4 addressed this.

P1, P2, Q1, Q2. Thank you for your answers and revisions.

I am satisfied with the revision wherein my concerns have been addressed, and think this manuscript does not require further review.

(Remarks on code availability)

Response for "Beyond representational alignment with brain-guided language models for robust reasoning"

Mingqing Xiao[1,4], Kai Du[2,3] and Zhouchen Lin[1,2*]

[1]State Key Lab of General AI, School of Intelligence Science and Technology, Peking University, Beijing, China.
[2]Institute for Artificial Intelligence, Peking University, Beijing, China.
[3]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China.
[4]Part of the revision work was done at Microsoft Research Asia.

*Corresponding author(s). E-mail(s): zlin@pku.edu.cn;
Contributing authors: mingqing_xiao@pku.edu.cn;
kai_du@tsinghua.edu.cn;

2 Response

We sincerely thank the Editor for handling our manuscript and the two reviewers for their careful evaluation, insightful comments, and constructive suggestions. In response, we have substantially revised both the manuscript and the Supplementary Information to address all concerns. Below, we first summarize the major revisions and then provide a point-by-point response to each comment.

1 Summary of Major Revision

1) Editorial compliance and code availability

In accordance with the journal's editorial requirements, we revised the title to remove punctuation, revised the abstract to have no more than 200 words, and removed terms such as "new", "novel", and "first" from the abstract and main text. We also updated the Code Availability statement to include a public repository and the corresponding reference DOI.

2) Updated neural predictivity analyses using functional localization

Following the suggestion from Reviewer 2 regarding the evaluation of neural predictivity, we re-ran the analyses using a functional localization procedure and updated corresponding results and descriptions. This provides a more scientifically valid estimate of neural predictivity, while leaving the main conclusions unchanged (Figure 2; Supplementary Information Figure S2; Methods; Supplementary Information Sections S3.1 and S7).

3) Validation on an independent external dataset

To assess generalizability beyond the original dataset in response to Reviewer 1 and Reviewer 2, we conducted new experiments on the HCP Relational Processing Task. As summarized in the main text and detailed in the Supplementary Information, these results show that the brain-guided framework is not limited to a single experimental design and can extend to a related

reasoning task with different stimulus modalities (Section 2.2; Supplementary Information Figure S14; Supplementary Information Section S13).

4) Evaluation of general capabilities and robustness

To address Reviewer 1's concerns about catastrophic forgetting and robustness, we evaluated all fine-tuned models on standard benchmarks, including MMLU, GSM8K, HumanEval, and MBPP. The results show that our method does not suffer from catastrophic forgetting of pre-existing general capabilities (Supplementary Information Table S1; Supplementary Information Section S11.5).

We also added analyses of subject-level robustness and individual differences (Supplementary Information Figure S11; Supplementary Information Sections S10.4 and S11.6), together with formal statistical tests for NARF+Label experiments based on two-sided paired t-tests across five runs (Figure 6; Supplementary Information Figure S8).

5) Clarification of intervention logic and expanded methodological discussion

In response to Reviewer 1, we clarified the conceptual rationale for applying intervention analyses only to problems that the model initially answers incorrectly in instance-specific intervention experiments (Supplementary Information Figure S13; Supplementary Information Section S12).

We also expanded the discussion of the method and the temporal resolution of fMRI with potential future directions (Supplementary Information Sections S5 and S6; Discussion).

6) Additional analysis results and discussions

In response to Reviewer 2, we added analyses of NARI (gen.) across intervention scales (Supplementary Information Figure S6; Supplementary

Information Section S10.3). In addition, we included new experiments combining NARF+Label with synthetic data, showing that neural guidance provides gains that are complementary to common techniques (Supplementary Information Figure S12; Supplementary Information Section S11.7). We also expanded the discussion of alignment scores (Supplementary Information Sections S7 and S11.3), and extended the results and discussions of the relation between subject-specific success rate and the human accuracy (Supplementary Information Section S10.2).

7) Presentation

Following the suggestions from Reviewer 1 and Reviewer 2, we improved the presentation by redesigning Figure 3 for improved readability (Figure 3) and standardizing the presentation of models and statistical tests (Figures 2, 4, 5, and 6; Supplementary Information Figures S5, S7, S8, and S9).

We believe that these revisions have substantially strengthened the rigor, clarity, and robustness of our work. We hope that our responses and the revised manuscript have adequately addressed the Editor's and Reviewers' concerns, and that you will find the revised manuscript suitable for publication in *Nature Machine Intelligence*.

2 Detailed Responses

2.1 To Editor

There are also some editorial points to take into account. Below, you will find our standard instructions to authors on revisions.

Response:

Thank you for the editorial guidance. We have made the following revision to fulfill the standard instructions:

- We have revised the title from "Beyond representational alignment: Brain-guided language models for robust reasoning" to "Beyond representational alignment with brain-guided language models for robust reasoning" so that no punctuation is used in the title.
- We have reivsed the abstrast to have no more than 200 words.
- We have revised the abstract and text to remove words of "new", "novel", and "first" for descriptions of the method and results.
- We have included the link to our code in the Code Availablility statement: https://github.com/pkuxmq/Brain-guided_LLM, and cited the corresponding DOI.

2.2 To Reviewer 1

We sincerely thank the reviewer for the thoughtful and constructive assessment of our work, and for appreciating our work's conceptual leap from correlational representational alignment to functional guidance. We have substantially strengthened the manuscript by adding new experiments, analyses, and explanations to address all concerns. Below we provide a detailed point-by-point response.

Major

1. General Capabilities and Risk of Catastrophic Forgetting

I note a critical lack of evidence regarding the impact of the proposed methods (NARI and NARF) on the model's broader capabilities. The study relies on a limited amount of task-specific fMRI data, raising serious concerns about overfitting. Catastrophic Forgetting: It is a common phenomenon in NeuroAI that aligning with biological signals or specialized fine-tuning can degrade performance on general tasks. It is unclear whether the observed improvements in deductive reasoning come at the cost of the model's pre-existing language understanding, generation, or commonsense reasoning abilities. Required Benchmarks: The authors must evaluate the NARF-tuned models on standard, broad-coverage benchmarks (e.g., MMLU, GSM8K, or general chat capabilities) to demonstrate that the model has not suffered from catastrophic forgetting. The effectiveness of the "regularization term" (Eq. 2) in preserving non-reasoning abilities must be empirically verified. Without this, it is difficult to rule out task-specific overfitting.

Response:

Thank you for this valuable suggestion. We agree that it is essential to verify whether neural-guided fine-tuning improves deductive reasoning without degrading broader model capabilities.

To address this concern, we supplement an evaluation of all fine-tuned models on standard benchmarks, including MMLU, GSM8K, HumanEval, and MBPP, following the OpenCompass [1] evaluation framework. The results (Tab. S1) show that performance on these general benchmarks remains essentially stable after fine-tuning, with only small fluctuations, and in several cases, performance even slightly improves. These results indicate that our method shows no evidence of catastrophic forgetting of the models' pre-existing general abilities.

Regarding the regularization term in Eq. (2), we would like to clarify that it is not designed to preserve non-reasoning abilities. Instead, this term is introduced only in the label-free NARF setting to regularize model representations for correct deductive reasoning problems, thereby reducing the systematic bias of NARF gradients from only incorrect problems. This role is specific to the reasoning task, which is necessary when labels are not used, does not involve non-reasoning tasks, and is not used in NARF+Label experiments.

The added benchmark results, along with the observed gains on generalization tasks (out-of-distribution syllogistic and transitive reasoning, propositional reasoning and natural language reasoning), support the conclusion that the fine-tuning is not task-specific overfitting.

Revision

We have added the evaluation results of general capabilities and the corresponding discussions in Supplementary Information Section S11.5 and Table S1. We quote the revision below:

"**S11.5 Evaluation of general capabilities**

To further evaluate whether fine-tuning with neural signals leads to catastrophic forgetting of general model capabilities, we evaluate all fine-tuned models on standard benchmarks, including MMLU, GSM8K, HumanEval, and

8 Response

Table S1 Evaluation of General Capabilities of Various Models after Fine-tuning.

Model	Method	Task Performance			
		MMLU	GSM8K	HumanEval	MBPP
Llama2-7B	Original	36.84	29.19	14.02	19.8
	NARF	37.18	29.04	14.02	19.2
	Label	41.08	28.13	14.63	20.2
	NARF+Label	37.39	28.43	13.41	19.2
Qwen2-1.5B	Original	55.54	61.11	43.90	30.8
	NARF	55.65	61.41	42.68	30.6
	Label	55.62	61.64	44.51	30.6
	NARF+Label	55.76	62.77	45.73	31.2
Phi4-mini	Original	74.47	83.02	78.05	55.4
	NARF	74.16	82.71	78.66	54.6
	Label	74.50	83.24	78.05	55.0
	NARF+Label	74.56	83.02	78.05	55.0
Llama3-8B	Original	67.60	79.23	58.54	58.0
	NARF	67.71	78.17	56.71	57.4
	Label	67.82	78.24	60.30	57.4
	NARF+Label	67.80	78.10	58.54	56.6
Qwen2-7B	Original	70.48	80.97	78.66	53.2
	NARF	70.36	80.36	76.83	53.2
	Label	70.44	81.12	78.66	53.0
	NARF+Label	70.33	79.83	77.44	55.4
Mistral-7B	Original	59.11	45.41	30.49	22.8
	NARF	59.36	45.34	34.76	25.2
	Label	59.46	46.47	30.49	23.6
	NARF+Label	59.46	46.10	31.10	23.4
Qwen3-4B	Original	75.79	86.81	79.27	65.4
	Label	76.06	87.19	79.27	65.4
	NARF+Label	75.84	86.43	78.05	65.6
Gemma2-9B	Original	74.74	79.23	23.78	59.8
	Label	74.25	79.45	25.61	59.2
	NARF+Label	74.33	80.29	25.61	59.4
Qwen2-72B	Original	82.34	91.66	83.54	63.6
	Label	82.50	91.28	82.32	67.6
	NARF+Label	82.29	91.96	82.93	65.8
Llama3.3-70B	Original	81.36	92.87	85.37	76.4
	Label	81.50	92.80	84.76	76.2
	NARF+Label	81.73	92.72	84.15	77.0

156 MBPP, following the OpenCompass [17] evaluation framework. The results

(Tab. S1) show that performance on these general benchmarks remains essentially stable after fine-tuning, with only small fluctuations, and in several cases, performance even slightly improves. These results indicate that our method does not suffer from catastrophic forgetting of general abilities."

2. Causal Interpretation and Intervention Logic

2.1 Intervention on Correct Samples*: The intervention is currently applied exclusively to cases where the model produces incorrect answers. This implies an assumption that correctly answered instances already possess "human-like" representations. This is neither justified nor validated. The authors must analyze the effects of neural-guided interventions on correctly answered samples. Does the intervention leave performance unchanged, or does it degrade correct reasoning trajectories? If correct responses are robust to intervention, the authors' claims are strengthened; if not, the causal link is weakened.*

Response:

Thank you for this important point. We agree that it is important to justify the influence on correctly answered instances and examine the causal link. We first clarify that our general-direction intervention is applied to both correctly and incorrectly answered questions, and then explain the conceptual rationale of only considering incorrect instances for causal examination in the instance-specific intervention setting.

Our general-direction intervention experiments (NARI (gen.), Fig. 4c,h,k) apply the brain-guided direction to the entire test set during inference, not only to initially incorrect examples. This evaluation therefore necessarily includes samples that the model already answers correctly. As shown in Fig. 4h,k, applying the brain-derived general direction improves overall test accuracy on new problems. While this does not isolate the initially correct subset, it

shows that applying the direction broadly does not destabilize model behaviour overall.

For the instance-specific intervention setting, however, it is more complicated and intervention on already-correct answers cannot serve as a meaningful causal test. The main reason is that humans are not perfect (e.g., no subject achieves 100% accuracy) and more "human-like" representations are not guaranteed to point to more correct direction for already-correct answers. As shown in Fig. S13 for a simplified illustration, suppose that we have a shared decision boundary between models and humans for correct and incorrect answers, for problems that both humans and models are correct, human representations can either be closer or farther to the boundary compared to model representations. Thus the direction to human representations can either strengthen the model decision or weaken the correct answer toward incorrect ones. Therefore, intervention on correct answers can in principle have both outcomes and cannot provide a causal examination of shared representational structure in the instance-specific setting.

By contrast, the model-incorrect / human-correct case is the most informative setting for testing the causal utility of neural guidance, as we can expect the direction toward human representations to guide incorrect points to correct ones if they have shared space. If steering toward the human-derived direction repairs a model error, this can inversely suggest a causal link. Additionally, for the other option, intervention on problems that model is correct while human is incorrect, it has trivial possibilities such as the language processing abilities of models are broken. Thus, intervention on incorrect answers toward correct ones is the only non-trivial case for our purpose.

For this reason, our instance-specific intervention analyses focus on problems that the model initially answers incorrectly, whereas our general-direction

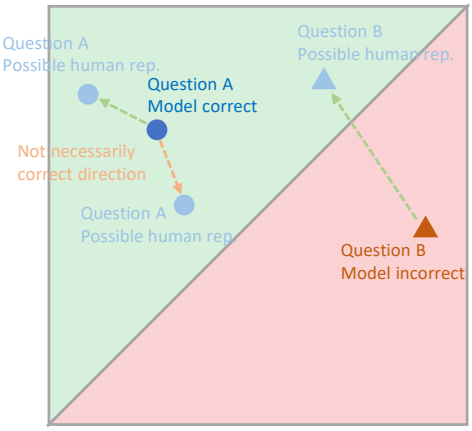

Fig. S13 Illustration of the experimental logic. We only intervene on problems that models are incorrect because for those that models are already correct, human representations do not necessarily provide correct direction.

intervention is evaluated on the full test set. We do not assume that "correctly answered instances already possess 'human-like' representation"; instead, we design the experiment setting based on the fact that human representations are not always perfect.

Revision

We have clarified this experimental logic in Supplementary Information Section S12 and Fig. S13. We quote the revision below:

"**S12 Extended explanation of experimental logic**

In this section, we further explain the logic of experiment design for only problems that the model is incorrect. The main reason is that humans are not perfect (e.g., no subject achieves 100% accuracy) and more "human-like" representations are not guaranteed to point to more correct direction for already-correct answers. As shown in Fig. S13 for a simplified illustration, suppose that we have a shared decision boundary between models and humans for correct and incorrect answers, for problems that both humans and models are

correct, human representations can either be closer or farther to the boundary compared to model representations. Thus the direction to human representations can either strengthen the model decision or weaken the correct answer toward incorrect ones. Therefore, intervention on correct answers can in principle have both outcomes and cannot inversely suggest whether human and model representations have shared subspace.

By contrast, the model-incorrect / human-correct case is the most informative setting for testing the causal utility of neural guidance, as we can expect the direction toward human representations to guide incorrect points to correct ones if they have shared space. If steering toward the human-derived direction repairs a model error, this can inversely suggest a shared subspace. Additionally, for the other option, intervention on problems that model is correct while human is incorrect, it has trivial possibilities such as the language processing abilities of models are broken. Thus, intervention on incorrect answers toward correct ones is the only non-trivial case for our purpose."

***2.2 Mechanism of Signal Effectiveness**: The reported noise ceilings of the reasoning-related brain regions are relatively low. A more explicit discussion is needed to explain how neural activations with limited signal-to-noise ratios can act as effective gradient directions or regularization signals.*

Response:

Thank you for this important point. We first clarify that the reported ceiling value is calculated based on cross-subject predictivity under the same regression pipeline, following the established practice in NeuroAI [2]. It is therefore not a direct estimate of the raw single-trial signal-to-noise ratio of single-trial fMRI responses, but rather an upper bound for neural predictivity. The results are similar to previous fMRI studies [2] and are not very low (as recognized by Reviewer 2).

More importantly, the effectiveness of our neural guidance does not rely on any single voxel/time point being highly reliable in isolation. Instead, the neural-guided direction is induced by the joint structure of model and brain representations across problems (Supplementary Information Section S5). The gradient depends on second-order terms such as $\mathbf{Y}_c\mathbf{Y}_c^\top$, and under this formulation, zero-mean noise in neural representations can be mitigated under expectation, because the expectation of the inner product between two zero-mean i.i.d. noise vectors is zero, whereas the shared problem-level structure remains and interacts with model representations to induce effective guidance directions.

Therefore, even when single-trial fMRI responses are imperfect, the method can still extract task-relevant structural information from the neural data and convert it into useful guidance signals.

Revision

We have added this discussion of robustness to neural noise in Supplementary Information Section S5. We quote the revision below:

"Additionally, this indicates that our neural-guided direction is, to some extent, robust to noises in neural signals for single voxel/time point, because it is induced by the joint structure of model and brain representations. The noise in neural representations can be mitigated in $\mathbf{Y}_c\mathbf{Y}_c^\top$ under expectation, while the remaining shared structure across problems can still interact with model representation to induce effective guidance directions."

***2.3 Intercept Ablation**: The finding that removing the intercept term b in ridge regression improves performance is counter-intuitive. Does this imply that the pattern (direction) of neural activity is informative, while the magnitude (baseline activation) introduces noise? A deeper discussion is warranted to clarify the mechanism of alignment.*

Response:

Thank you for this insightful question. We clarify the mechanism more explicitly. In our method, we compute intervention/fine-tuning directions from the gradient of a similarity objective defined on the transformed representation. When the intercept is included, the mapped prediction is $\mathbf{W}\mathbf{x}+\mathbf{b}$; when it is excluded, it is $\mathbf{W}\mathbf{x}$. Excluding the intercept term in the objective calculation induces an additional term in the direction (gradient) compared with the intercept-inclusive form, which represents systematic shift between model's and human's mean representations adjusted by the interactive structure between model and human representations. For example, taking Sim as the negative of MSE, for the intercept-inclusive form $S_{w/\ icpt} = \mathrm{Sim}(\mathbf{W}\mathbf{x}+\mathbf{b}, \mathbf{y})$, the gradient is $\nabla_{\mathbf{x}} S_{w/\ icpt} = -\mathbf{W}^\top\mathbf{W}\mathbf{x} - \mathbf{W}^\top\mathbf{b} + \mathbf{W}^\top\mathbf{y}$, while the gradient for the intercept-exclusive form is $\nabla_{\mathbf{x}} S_{w/o\ icpt} = -\mathbf{W}^\top\mathbf{W}\mathbf{x} + \mathbf{W}^\top\mathbf{y}$, where $\mathbf{W} = \mathbf{Y}_c^\top\mathbf{X}_c\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}$ is induced by the structures from centered model and human representations and $\mathbf{b} = \overline{\mathbf{y}} - \mathbf{W}\overline{\mathbf{x}}$ ($\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are sample means) is a systematic shift between model and human representations (Supplementary Information Section S5). This means $\nabla_{\mathbf{x}} S_{w/o\ icpt} - \nabla_{\mathbf{x}} S_{w/\ icpt} = \mathbf{W}^\top\mathbf{b}$ so that the gradient ascent direction additionally incorporate a direction induced by the systematic shift.

This indicates that not only structures from centered representations (covariance) are important but also systematic shift between mean representations can contribute to guiding directions when modulated by $\mathbf{W}$ from the joint structures. It implies that both the pattern of neural activity and magnitude of activation are informative.

Revision

We have added this expanded discussion in Supplementary Information Section S5. We quote the revision below:

"Note that we calculate the gradients using the objective $S = \text{Sim}(\mathbf{Wx}, \mathbf{y})$ rather than the intercept-inclusive form $S_{w/\ icpt} = \text{Sim}(\mathbf{Wx} + \mathbf{b}, \mathbf{y})$, which empirical results demonstrate yields superior performance (Sec. S10.2). Different from the intercept-inclusive form that aims to directly encourage the similarity between centered model and human representations, this design induce an an additional term $\mathbf{W}^\top\mathbf{b}$, representing a systematic shift between model and human representations (*i.e.*, $\mathbf{b}$) adjusted by the interactive structure between model and human representations (*i.e.*, $\mathbf{W}$): for example, taking Sim as the negative of MSE, the gradient for the intercept-inclusive form is $\nabla_{\mathbf{x}} S_{w/\ icpt} = -\mathbf{W}^\top\mathbf{Wx} - \mathbf{W}^\top\mathbf{b} + \mathbf{W}^\top\mathbf{y}$, while the gradient for the intercept-exclusive form is $\nabla_{\mathbf{x}} S = -\mathbf{W}^\top\mathbf{Wx} + \mathbf{W}^\top\mathbf{y}$, meaning $\nabla_{\mathbf{x}} S - \nabla_{\mathbf{x}} S_{w/\ icpt} = \mathbf{W}^\top\mathbf{b}$ so that the gradient ascent direction additionally incorporate a direction induced by the systematic shift. As shown in Sec. S10.2, neither component alone (systematic shift alone or without this term) achieves optimal results. This indicates that not only structures from centered representations are important but also systematic shift between mean representations can contribute to guiding directions when modulated by $\mathbf{W}$ from the joint structures. Our method thus provides a new way to induce directions for improving model representations, instead of directly maximizing similarity scores that usually measure the similarity for centered data and have been shown not to guarantee encoding task-relevant information [10]."

3. Data Scale, Robustness, and Cross-Subject Generalization

3.1 *The study utilizes a very small fMRI dataset (<70 unique problems from 10 subjects) to train models evaluated on >50,000 synthetic test problems. This mismatch raises concerns about the universality of the extracted signals.*

Response:

Thank you for this important question. We agree that the contrast between a small fMRI dataset and a large synthetic test set requires clarification.

In our study, these two components serve different purposes. The fMRI dataset is used to derive the neural guidance signal, whereas the large-scale synthetic test set is used to evaluate whether that guidance generalizes beyond the original neural samples. The large number of test problems is our evaluation design for testing the "universality" of extracted signals, that is, generalizing to new reasoning problems instead of overfitting to or memorizing the specific problems in the fMRI dataset. The large-scale evaluation minimizes random variance, thereby enabling a more reliable assessment of whether the signal learned from a limited number of neural examples reflects a shared reasoning-related component that generalizes to held-out problems.

Therefore, our large synthetic test set represents a more rigorous test of universality rather than a methodological mismatch.

***3.2 Leave-One-Subject-Out (LOSO) Analysis**: The current method aggregates signals from all 10 subjects. To prove that the method captures a universal reasoning mechanism rather than overfitting to specific subjects, the authors must perform a LOSO analysis (train on N-1 subjects, evaluate on the held-out subject).*

Response:

Thank you for this important suggestion. We agree that it is necessary to distinguish a shared reasoning-related signal from subject-specific overfitting.

Although a literal leave-one-subject-out (LOSO) analysis is not directly applicable to our experiments, because our experiments do not predict or evaluate subjects' neural responses at test time, we agree with the reviewer's

underlying concern: the key question is whether the extracted guidance generalizes beyond the specific subjects and stimuli used to derive it.

Our primary evaluation is designed to address this point. Specifically, we use neural signals during training for guidance construction, but evaluate the resulting models on newly generated reasoning problems under multiple settings, including out-of-distribution test conditions. This design asks whether the learned guidance transfers to held-out reasoning problems, rather than remaining tied to the original subject-specific signals or stimulus set. In this sense, the observed improvements on new questions argue against simple overfitting to the specific subjects or problems in the fMRI dataset.

As supporting evidence, we also conducted additional subject-robustness analyses (see our response to Point 9 below and Fig. S11). These analyses show that the performance improvement remains generally stable when the number and composition of subjects are varied. Although these analyses are not a literal LOSO analysis, they provide further evidence that the method is not driven only by specific subjects.

Taken together, our testing results of generalization and the added subject-robustness analyses support the conclusion that the extracted neural guidance captures a shared reasoning-related component rather than overfitting to specific subjects.

***3.3 External Dataset Validation**: Relying on a single dataset (Lytle et al., 2020) limits the robustness of the conclusions. Validation on an independent dataset (e.g., HCP Relational Processing Task or Pereira et al., 2018) is strongly recommended to demonstrate that the pipeline is not an artifact of one specific experimental design.*

Response:

Thank you for this valuable suggestion. We agree that validation on an independent external dataset is important to demonstrate that the proposed pipeline is not an artifact of one specific experimental design.

To address this concern, we add new experiments on an external dataset, HCP Relational Processing Task from the Human Connectome Project [3], as shown in Fig. S14. The HCP Relational Processing Task probes relational reasoning by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top images differ in (Fig. S14b). We adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a). Such different modalities extend our experimental settings to evaluate another abstract relational reasoning ability with different stimulus modalities. This dataset differs from our original deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

We conduct both intervention and fine-tuning experiments on this task and dataset using five models from different families (Phi4-mini, Qwen2-7B, Llama3-8B, Gemma2-9B, Mistral-7B), and the implementation details are presented in Supplementary Information Section S13 (quoted below). As shown in Fig. S14c, human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges. While the overall intervention success rate on this dataset (around 80%) is lower than the original deductive reasoning setting (100%), it is likely due to an increased difficulty introduced by the modality shift, which reflects a boundary condition while remaining

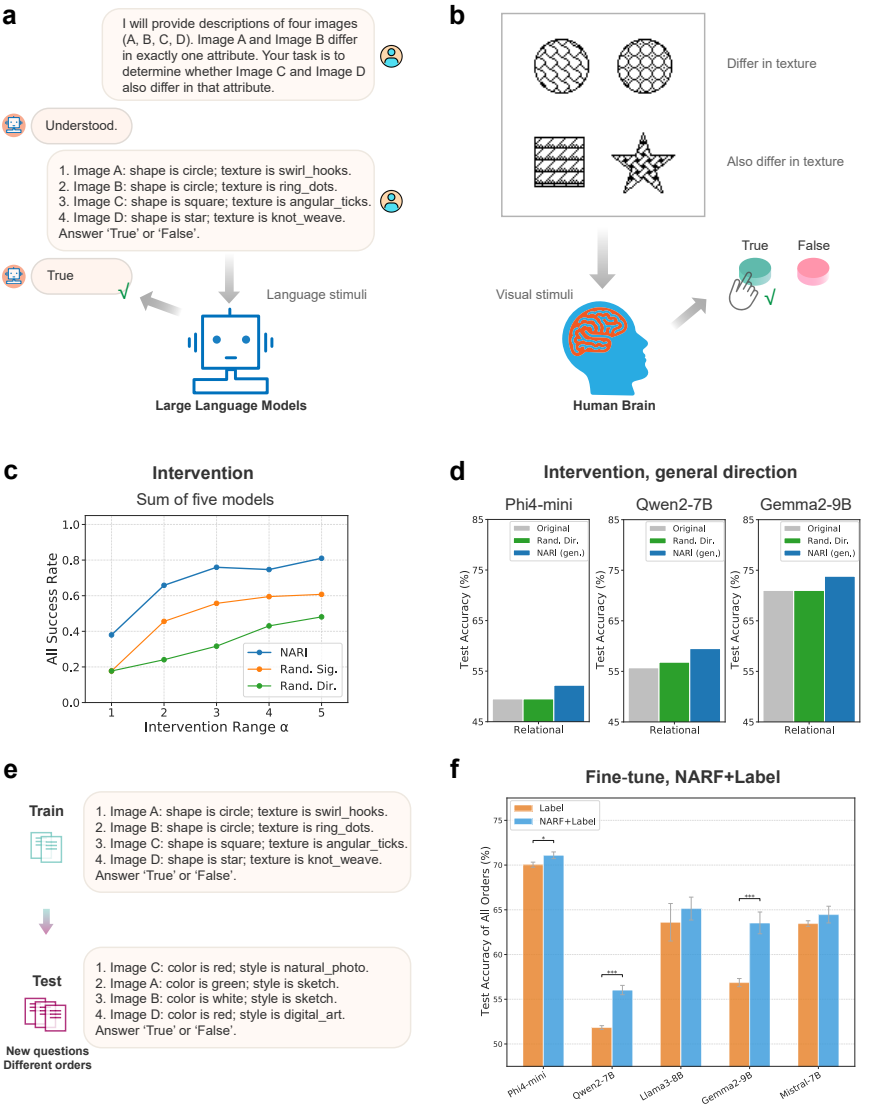

Fig. S14 Results on the Relational Processing Task from the HCP Dataset. (a-b) Illustration of the task for LLMs and humans. The task probes relational processing by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top image differ in. LLMs process problems through chat-formatted prompts with language-based descriptions of image attributes, generating direct linguistic responses. Human participants view two pairs of images (top and bottom), responding via button press. (c) Summary results of five LLM models for intervention on their incorrect problems. NARI achieves much higher success rates than baselines of random signals and the max-integration of random directions. (d) Results of the general direction on the new test data. (e) Illustration of the training and testing settings. (f) Testing accuracies of various models for language-label-only and NARF+Label fine-tuning.

effective. In addition, Fig. S14d further shows that the general direction is possible to transfer to new, out-of-set problems. Finally, in a similar fine-tuning setting with evaluation on new questions under different orders (Fig. S14e), NARF complements language supervision with consistent gains over language supervision alone (Fig. S14f).

These results on an independent dataset show that the pipeline is not an artifact of one specific experimental design. Moreover, they show that our framework can extend to another reasoning domain with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data.

Revision

We have added these new results, discussions, and details in Supplementary Information Section S13 and Fig. S14, and briefly introduced this extension in the main text (Section 2.2). We quote the revision below.

The revision in Section 2.2 is:

"Extension to other neural datasets We further validate our method on another independent dataset, HCP Relational Processing Task from the Human Connectome Project [48] (Sec. S13, Fig. S14). This task probes relational reasoning by comparing relations from two pairs of images, and we adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a-b). This dataset differs from the deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

The results show that human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges (Fig. S14c), and the general direction

of NARI is possible to transfer to new, out-of-set problems (Fig. S14d). At the same time, NARF complements language supervision with consistent gains (Fig. S14f). Full results and details can be found in Sec. S13.

The results show that our pipeline is not limited to one specific experimental design. Meanwhile, they show that our method can extend to other related reasoning tasks with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data."

The revision in Supplementary Information Section S13 is:

"**S13 Extension to the HCP relational processing dataset**

In this section, to further validate our method on other independent datasets, we extend our experiments to HCP Relational Processing Task from the Human Connectome Project [18] (Fig. S14). The HCP Relational Processing Task probes relational reasoning by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top images differ in (Fig. S14b). We adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a). Such different modalities extend our experimental settings to evaluate a different abstract relational processing ability with different stimulus modalities. This dataset differs from our original deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

Details of the fMRI dataset and preprocessing The dataset is derived from the relational processing task in the Human Connectome Project (publicly available on www.humanconnectome.org), which originally contains more than 1,000 human subjects. We randomly select 20 subjects (IDs: 100610, 103111, 139637, 149236, 162733, 169444, 200008, 200513, 211619, 308129, 421226, 441939, 529549, 567759, 611938, 727553, 771354, 793465, 872158,

984472), whose behavioural accuracy ranges from 70% to 100% for this task. The dataset probes higher-order relational reasoning using visual stimuli that vary in shape and texture. It includes relational and matching (the control condition) blocks across two runs for each subject. For the relational block, each trial presents two pairs of images, where the top images differ in exactly one attribute (shape or texture). Participants are required to judge whether the bottom images also differ in that attribute and respond via button press. Each subject has a total of 24 relational trials (2 runs, 3 relational blocks each run, and 4 trials each block).

While the dataset follows a block-designed task-fMRI paradigm, we extract the stimuli for each trial from the accompanying description file of experimental designs for each subject (which describes the file name of each stimuli) and the publicly available stimuli files (from the HCP_TFMRI_scripts). As the stimuli files are images, we manually annotate the shape and texture attributes of each image and convert them into language descriptions. There are 96 stimuli files in the scripts, while we find that only 47 stimuli files appear in fMRI experiments of the 20 subjects. There are total six possible shapes and six possible textures for images.

We use the fMRI data preprocessed by the Human Connectome Project, and estimate the response to each trial with GLMSingle. We use the average response time as the response time in GLMSingle. For the design matrix, we consider up to 3 conditions (used for cross-validation in GLMSingle): correct answer for questions with "yes" answer, correct answer for questions with "no" answer, and wrong answer, and extract single-trial responses for each problem with the design matrix.

We extract voxels from the estimated response based on locations of MD regions and the subject-specific functional localization process by contrasting

relational blocks over matching blocks. We first use SPM12 to get the t-values of all voxels considering the contrast, and then select 10% of the most responsive voxels within the MD mask. These voxels are considered as vectors of brain representations for each human subject.

Details of model prompting We convert image stimuli to language descriptions for LLMs to process. Specifically, we convert the six possible shapes into: *circle, cross, triangle, square, star, hexagon*, and convert the six possible textures into: *ring_dots, angular_ticks, zigzag_dashes, solid_blocks, knot_weave, swirl_hooks*. The task prompt for the model is:

"*You are an expert in relational reasoning. I will provide descriptions of four images (A, B, C, D), which describe two attributes of each image: shape and texture. Image A and Image B differ in exactly one attribute (either shape or texture). Your task is to determine whether Image C and Image D also differ in that attribute. For example, if Image A and Image B have different shapes, then you should determine if Image C and Image D also have different shapes. Or if Image A and Image B have different textures, then you should determine if Image C and Image D also have different textures. If the statement is true, you should answer 'True'; otherwise answer 'False'. Only answer 'True' or 'False' at the beginning of the response.*"

After the model's round saying "*Understood. I will analyze the attributes and perform relational reasoning to judge the statement. I will only answer 'True' or 'False' at the beginning of the response.*", we present the descriptions as follows:

"*1. Image A: shape is circle, texture is solid_blocks.\n 2. Image B: shape is circle, texture is ring_dots.\n 3. Image C: shape is triangle, texture is zigzag_dashes.\n 4. Image D: shape is hexagon, texture is zigzag_dashes.\n For the attribute that Image A and Image B differ in (either shape or texture),*

do Image C and Image D also differ in that attribute? Only answer 'True' or 'False' at the beginning of your response."

Then we extract the first token of the model's response.

Generated test dataset For testing of LLMs, we generate new questions with different descriptions of attribute types and attributes. Specifically, we replace "shape" and "texture" by two random attribute type in *[color, quality, geometry, style]*, and each of them has the following possible attributes: *color: [red, green, blue, yellow, black, white]; quality: [normal, noisy, blurred, underexposed, overexposed, compressed]; geometry: [rotated, translated, zoomed_in, zoomed_out, flipped, sheared; style: [sketch, watercolor, oil_painting, digital_art, cartoon, natural_photo]*. There are four conditions: true/false $\times$ first/second attribute is different. For the test data, we generate 100 problems for each condition, leading to 400 problems; for the validation data for fine-tuning, we generate 20 problems for each condition, leading to 80 problems. When evaluating fine-tuned models, we also enumerate all possible permutations for the order of descriptions of four images (Fig. S14e), with finally $400 \times 24 = 9600$ test data.

Implementation details of NARI and NARF We leverage the similar experimental settings as the deductive reasoning task. For NARI, the maximum gradient descent steps are set as 50. For the general direction of NARI, we take the average of successful intervention directions with $\alpha = 3$ (for Phi4-mini, we take $\alpha = 5$), normalize the direction for each layer to the standard deviation scale, and the scale γ is searched between 0.5 and 10 with the interval 0.5. For the combination of NARF and language labels, we use all 96 questions from the HCP stimuli files with only part of them having fMRI signals. We select 10 subjects for fine-tuning based on the intervention success rate, and only perform guidance for questions that the model can be

intervened to the correct answer. We start training of NARF+Label from the language-label-trained models, except for Phi4-mini that we find starting from the original model has better performance. λ_{narf} is set as 0.1 for Phi4-mini and Llama3-8B, 0.05 for Qwen2-7B, and 0.01 for Mistral-7B and Gemma2-9B. We also run each experiment for 5 times with the same random seeds: 0, 1, 42, 1000, 2024, and report statistical significance based on two-sided paired t-test among the 5 runs.

Results We conduct intervention and fine-tuning experiments on this task and dataset using five models from different families (Phi4-mini, Qwen2-7B, Llama3-8B, Gemma2-9B, Mistral-7B). As shown in Fig. S14c, human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges. While the overall intervention success rate on this dataset (around 80%) is lower than the original deductive reasoning setting (100%), it is likely due to an increased difficulty introduced by the modality shift, which reflects a boundary condition while remaining effective. In addition, Fig. S14d further shows that the general direction is possible to transfer to new, out-of-set problems. Finally, in a similar fine-tuning setting with evaluation on new questions under different orders (Fig. S14e), NARF complements language supervision with consistent gains over language supervision alone (Fig. S14f).

These results on an independent dataset show that our pipeline is not limited to one specific experimental design. Meanwhile, they show that our method can extend to other tasks with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data."

4. Domain Shift and Failure Analysis

4.1 Pseudowords vs. Natural Language*: The fMRI task uses pseudowords*

to isolate logic, but LLMs are pre-trained on natural language. Please discuss whether this distribution shift affects the efficiency of the guidance. Does the method work equally well on natural language reasoning datasets (e.g., FOLIO) compared to the synthetic pseudoword test set?

Response:

Thank you for this insightful question. We agree that the potential distribution shift between pseudoword-based neural stimuli and natural-language-pretrained LLMs deserves explicit discussion.

The use of pseudowords in the deductive reasoning stimuli is intentional, because it helps minimize semantic confounds and potential data contamination from language pretraining, thereby isolating deductive reasoning ability more cleanly from lexical or world-knowledge effects. Similar synthetic pseudowords have also been used in prior evaluations of LLMs [4], showing that LLMs are not limited to natural language. This distribution shift is part of the design to better probe true reasoning ability.

We do not observe evidence that this design has significant affects on guidance. Meanwhile, for natural language reasoning, we find that NARF improves generalization to the FOLIO dataset (Fig. 6b), indicating that representational guidance learned from pseudowords can also benefit reasoning expressed in natural language.

As for guidance with natural language reasoning, due to the lack of neural data for natural language reasoning, we are unable to evaluate the full pipeline in this setting. We hope our work can inspire and encourage the study and release of more task-specific neural data for broader future investigation.

4.2 Granular Failure Analysis*: While the authors analyze failure cases based on subject performance, a more granular analysis is needed. Are there*

specific linguistic characteristics (e.g., negation, sentence complexity) or logical structures where the fMRI signal consistently fails to steer the model?

Response:

Thank you for this important point. We agree that failure analysis is important for understanding the strengths and limits of a method, and we provide more analyses.

For the original deductive reasoning questions from the fMRI dataset, NARI ultimately achieves a 100% success rate for intervention. Therefore, within this dataset, we do not observe specific linguistic characteristics or logical structure that the fMRI signal consistently fails to steer.

For the newly generated test questions, we performed a more granular analysis by reasoning subtype, covering different linguistic characteristic (affirm/negation), different complexity (requiring two of three or all three premises for the logic), and different reasoning types (syllositic/transitive). Fig. S7 (Supplementary Information Section S11.1) presents the performance stratified by reasoning subtypes for models before and after various fine-tuning settings. While the magnitude of improvement varies across models and subtypes, we do not observe specific characteristics for which fMRI signals consistently fails to improve models.

Taken together, our current results indicate no evidence for consistent failure characteristics.

Minor

5. Computational Cost*: Please clarify the additional computational latency introduced by the NARI (inference-time intervention) method. Is it feasible for real-time applications?*

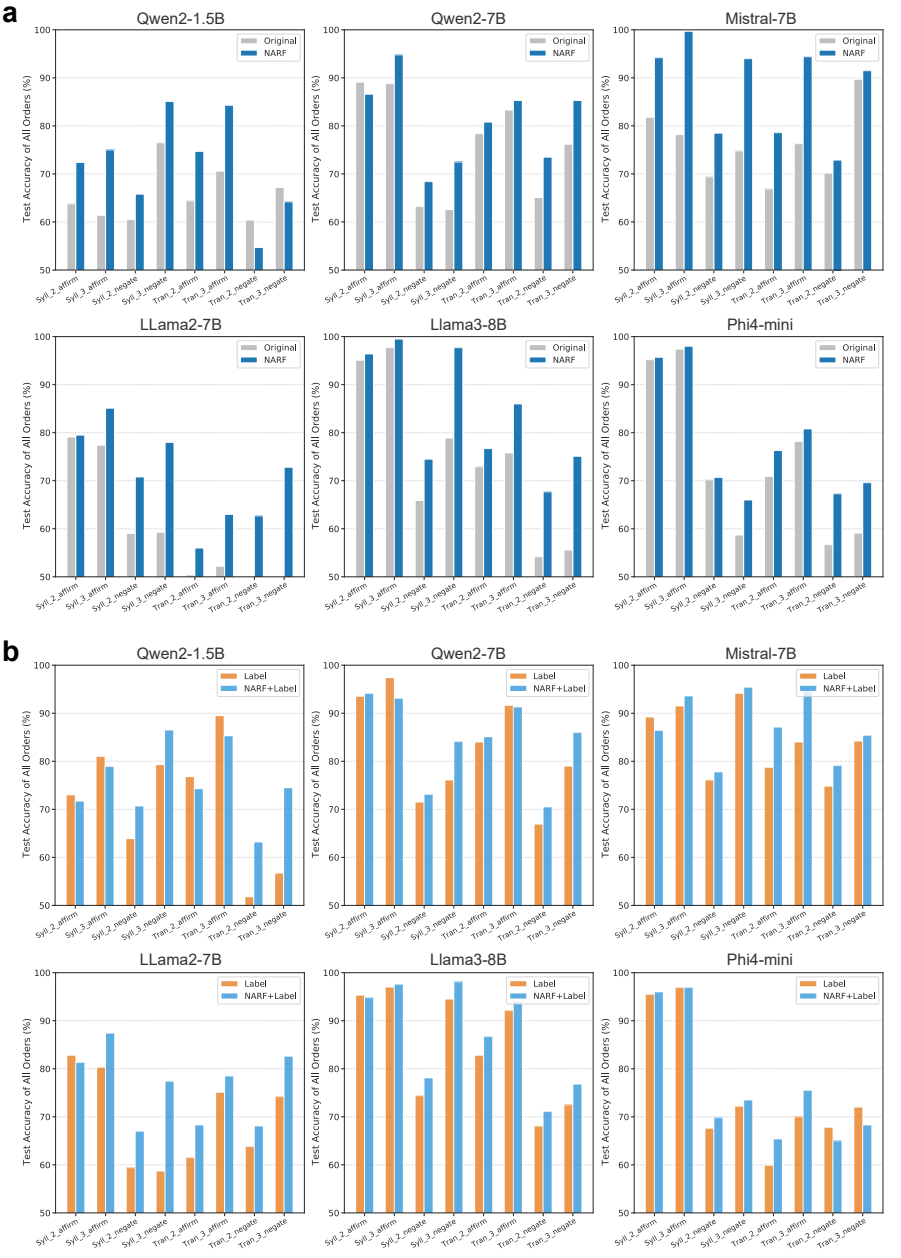

Fig. S7 More Results on Performance across Reasoning Subtypes after Fine-tuning. (a) Comparison results between the original and Neural Activation guided Representation Fine-tuning (NARF) fine-tuned models. (b) Comparison results between the language-label-only fine-tuned and NARF+Label fine-tuned models.

Response:

Thank you for this important question. We provide more detailed clarification. For the general direction of NARI (gen.), the intervention is a simple additive steering vector applied to intermediate representations in middle layers using precomputed directions. This introduces negligible overhead beyond the standard forward pass, and is feasible for real-time applications as other steering methods [5]. For the per-instance direction calculation of NARI, we apply projected gradient ascent to modify representations based on the objective and verify the direction every 5 steps until success or reach maximum steps (set as 200 or 50). The gradient ascent for representations only have small costs because it mainly involves one linear transformation for the objective calculation, while the verification process requires inferring the whole model with intervention directions. Therefore, the costs mainly comes from the verification process, which can take between 1-40 (or 1-10) times of LLM inference time for each fMRI data depending on the difficulty of successful intervention.

Revision

We have added this discussion of the computational cost of NARI in Supplementary Information Section S6. We quote the revision below:

"Considering computational costs, for the general direction of NARI (gen.), the intervention is a simple additive steering vector applied to intermediate representations in middle layers using precomputed directions. This introduces negligible overhead beyond the standard forward pass, and is feasible for real-time applications as other steering methods [11]. For the per-instance direction calculation of NARI, the gradient ascent for representations only have small costs because it mainly involves one linear transformation for the objective calculation, while the verification process requires inferring the whole model with intervention directions. Therefore, the costs mainly comes from the verification

process, which can take between 1-40 (or 1-10) times of LLM inference time for each fMRI data depending on the difficulty of successful intervention."

***6. Visualization**: Figure 3 (Methodology) is somewhat cluttered. The distinction between the "Intervention" path and the "Fine-tuning" path could be made visually distinct (e.g., different color coding for data flows).*

Response:

Thank you for this valuable suggestion. We have revised Fig. 3 to improve readability by visually separating the Intervention and Fine-tuning parts with different background colors. We quote the figure in Fig. 3.

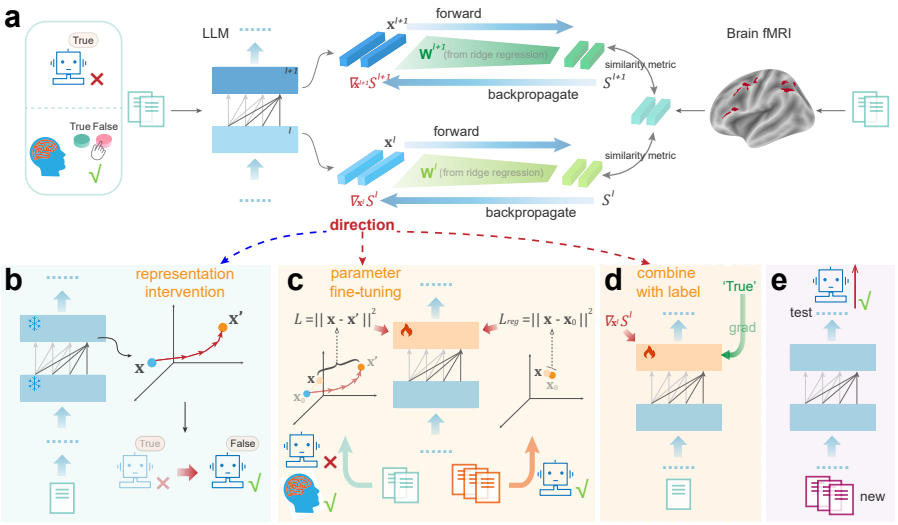

Fig. 3 Illustration of the Methods to Steer Large Language Models (LLMs) with Human Neural Activations. (a) We induce directions at the model representation space of each layer for problems with incorrect responses. We calculate gradients of model representations $\mathbf{x}$ to maximize a similarity metric between functional magnetic resonance imaging (fMRI) signals and $\mathbf{x}$ transformed to the fMRI space via weights from ridge regression, which encodes the structures of model and fMRI representations. (b) We apply these directions to modify representations through inference-time intervention without parameter updates, correcting model outputs. (c) The directions that induce successful intervention are used to fine-tune model parameters, injecting representational knowledge into the parameters. (d) The neural-derived directions complement output language supervision during fine-tuning through the combination of objectives. (e) The models improve test performance on new deductive reasoning problems.

***7. Terminology**: The terms "Deductive Reasoning Network" and "Logic Network" appear to be used interchangeably. Consistent terminology would improve readability.*

Response:

Thank you for this suggestion. We have thoroughly check the manuscript and ensure that no "Logic Network" is used, with consistent terms of "deductive reasoning network" (or shorten for "reasoning network").

***8. Statistical Significance**: In the OOD generalization results (Fig. 6), please ensure that statistical significance tests are reported for the performance gaps between "Language SFT" and "NARF".*

Response:

Thank you for this valuable suggestion. Here we additionally conduct more experiment runs for each model under each fine-tuning setting with the same random seeds, and updated Fig. 6 in the main text as well as Fig. S8 in the Supplementary Information to report statistical significance based on two-sided paired t-test over 5 runs. This enhances the statistical rigor of our results.

Revision

We have updated Fig. 6 and Fig. S8, as well as the descriptions of the results and experiment details in the main text (Sections 2.2.3, 4.4, and 4.6) and Supplementary Information (Section S6). We quote the revision below.

The details in Sections 4.4, 4.6 and Supplementary Information Section S6 are:

"For fine-tuning experiments, statistical testing also employs two-sided paired t-tests comparing NARF+Label and Label under the same random seeds."

"We run each experiment for 5 times under the same random seeds for statistical significance test. We apply Low-Rank Adaptation (LoRA) [65] to

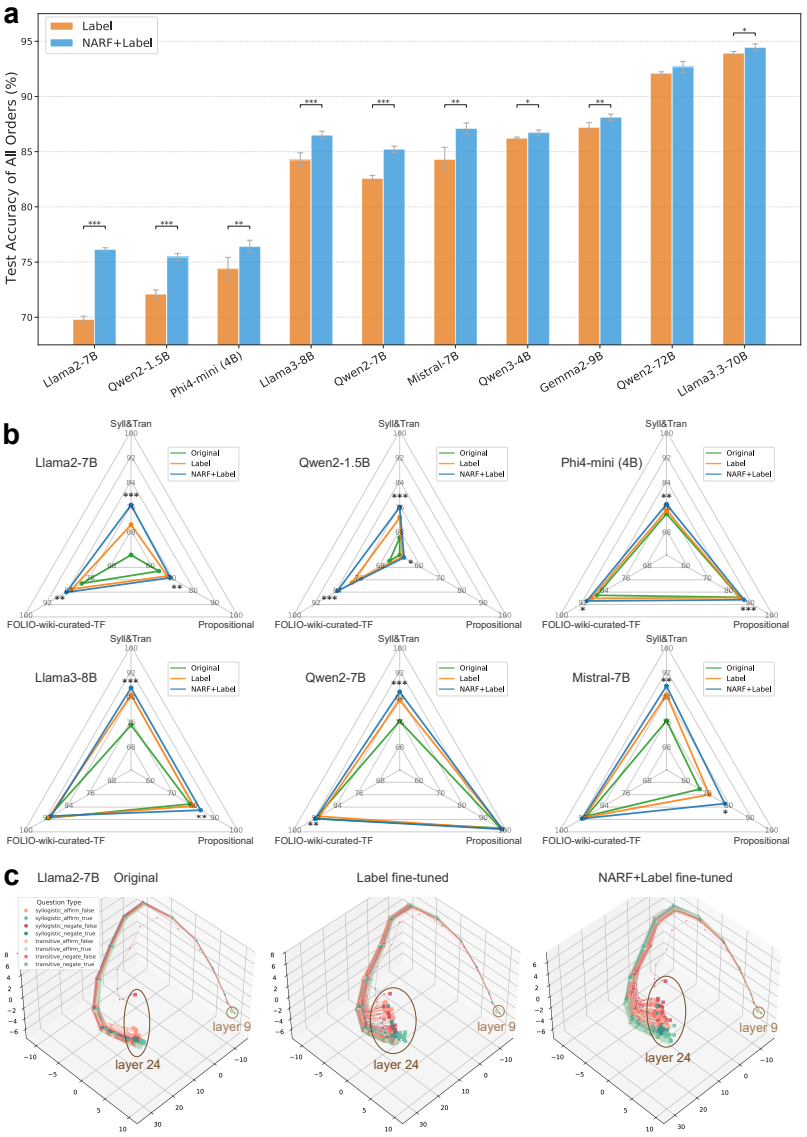

Fig. 6 Results of Combining NARF with Language Supervision. Statistical tests are reported based on two-sided paired t-test over five runs of experiments under same random seeds. (a) Testing accuracies (mean±std) of various models (1.5B–70B parameters) for language-label-only and hybrid (NARF+language labels) fine-tuning, ordered based on models' original test accuracy. (b) Testing performance (mean) comparison on diverse reasoning data. Improvements on syllogistic and transitive problems generalize to propositional reasoning and the FOLIO dataset with first-order logic in natural language. (c) Visualization of the layer-wise reasoning trajectories (layer outputs from the 1/4-th layer to the 3/4-th layer) in the representation space shows enhanced separation between paths for correct and incorrect logics after hybrid fine-tuning compared to the original and label-only fine-tuned models.

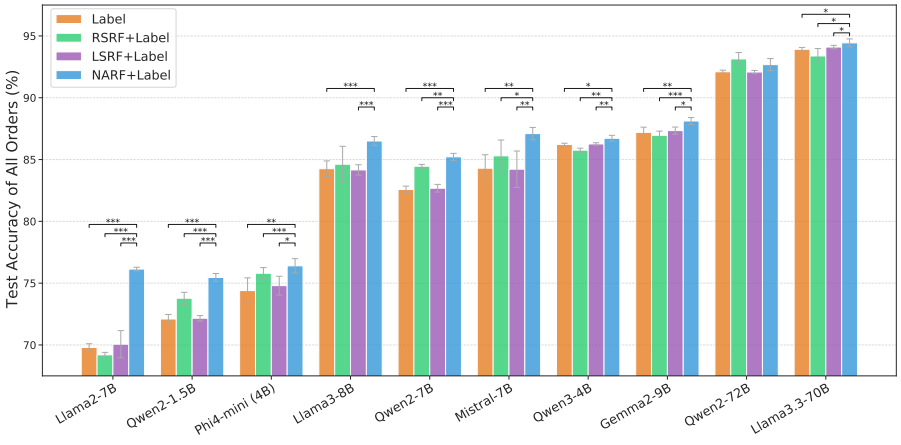

Fig. S8 Ablation Analysis of NARF when Combined with Language Supervision. RSRF replaces fMRI signals with random data, which eliminates the effect of human neural activations and is only induced by the representational structures of models. LSRF replaces fMRI signals with one-hot encodings of label signals, representing the joint effect by the representational structures of models and labels. Results are ordered based on the models' original test accuracy.

the parameters of attention modules in the same middle layers and utilize the Adam optimizer to update them (learning rate $1e-4$, no weight decay), with accumulated gradients from all training problems. Models are trained for 100 epochs and validated every epoch (for Qwen2-72B-Instruct and Llama3.3-70B-Instruct, we train the models for 50 epochs). λ_{narf} is set as 0.1 for most models and 0.01 for Qwen2-1.5B-Instruct. For the baseline of language label fine-tuning alone, it keeps the same setting and can be viewed as taking $\lambda_{narf} = 0$."

"We run each experiment for 5 times with the same random seeds: 0, 1, 42, 1000, 2024, and report statistical significance based on two-sided paired t-test among the 5 runs."

The updated description of the results in Section 2.2.3 is:

"The results show that augmenting language supervision with NARF yields consistent, statistically significant gains (Fig. 6a), with 2.2% average accuracy gain in average (range 0.5–6.4%) on test problems with all premise-order permutations."

"For instance, on propositional problems, Mistral-7B-Instruct is improve by 13.2% considering the max performance under five runs (83.7%) over the max performance of language-only fine-tuning (70.5%)."

***9. Regarding Individual Differences and Insufficient Robustness Analysis**: The paper notes significant variability in intervention success rates across subjects (Fig. S5), which correlates with subjects' behavioural accuracy. This suggests that the effectiveness of neural guidance may be highly dependent on the quality of individual neural representations. We recommend a further analysis to clarify whether the method is primarily effective only for signals from "high-performing subjects". Crucially, it should be investigated whether signals from subjects with lower alignment or performance could introduce noise and potentially degrade model performance. Addressing this is essential for assessing the generalizability and practical robustness of the proposed guidance approach.*

Response:

Thank you for this valuable suggestion. Here we supplement analysis experiments of NARI (gen.) and NARF+Label with different number of subjects. We vary the number of subjects from 1 to 10 by gradually including subjects with two directions: from high-performing to low-performing subjects and vice versa. As shown in Fig. S11a-b, increasing the number of subjects gradually improves the performance of NARI (gen.), where high-performing subjects generally have slightly better performance, while including low-performing subjects does not degrade the performance. As shown in Fig. S11c, it is also

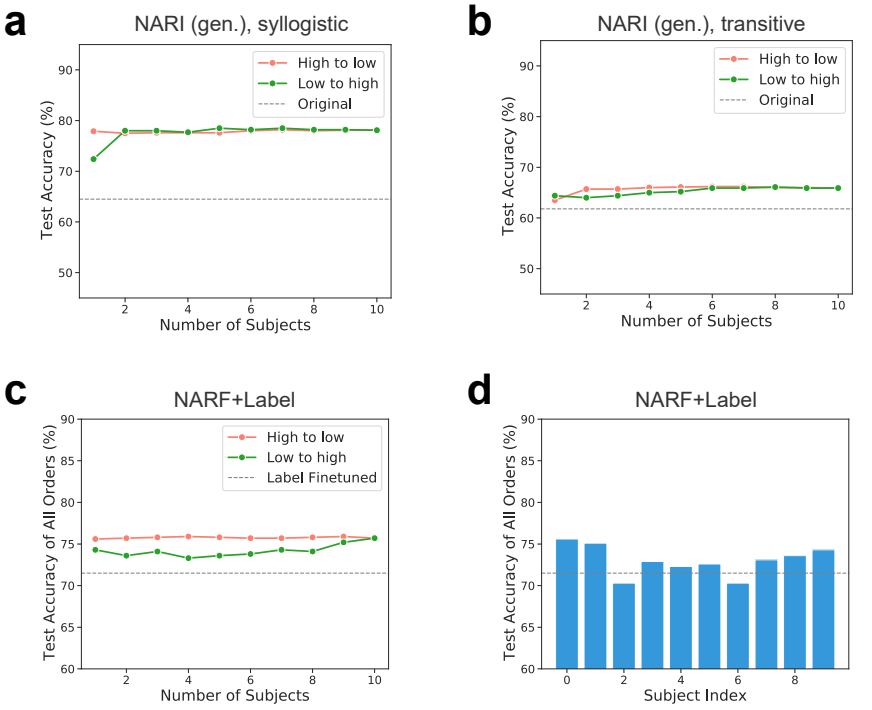

Fig. S11 Analysis Results of the Influence of Different Human Subjects on NARI (gen.) and NARF+Label. (a-b) Test performance of NARI (gen.) on syllogistic and transitive problems with varying number of subjects in two directions: from high-performing to low-performing subjects and vice versa. (c) Test performance of NARF+Label with varying number of subjects in two directions: from high-performing to low-performing subjects and vice versa. (d) Test performance of NARF+Label for each subject.

similar for NARF+Label that including high-performing subjects have better performance while including low-performing subjects given high-performing subjects does not degrade the performance. Fig. S11d further demonstrates the performance for each subject, showing that the method is not only effective for high-performing subjects, and while not all subjects have uniform gains, including these subjects does not degrade model performance. There results verify the robustness of our method.

Revision

We have added these robustness analysis results and discussions in Supplementary Information Sections S10.4, S11.6, and Fig. S11. We quote the revision below:

"**S10.4 Analysis results of subject-wise robustness**

We further analyze the robustness of human subjects. We conduct experiments of NARI (gen.) with different number of subjects by varying the number of subjects from 1 to 10 through gradual inclusion of subjects with two directions: from high-performing to low-performing subjects and vice versa. As shown in Fig. S11a-b, increasing the number of subjects gradually improves the performance of NARI (gen.), where high-performing subjects generally have slightly better performance, while including low-performing subjects does not degrade the performance. There results verify the robustness of the our method."

"**S11.6 Analysis results of subject-wise robustness**

We also analyze the robustness of human subjects for NARF+Label. We vary the number of subjects from 1 to 10 through gradual inclusion of subjects with two directions: from high-performing to low-performing subjects and vice versa. As shown in Fig. S11c, it is also similar for NARF+Label that including high-performing subjects have better performance while including low-performing subjects given high-performing subjects does not degrade the performance. Fig. S11d further demonstrates the performance for each subject, showing that the method is not only effective for high-performing subjects, and while not all subjects have uniform gains, including these subjects does not degrade model performance. There results verify the robustness of the our method."

10. Regarding Temporal Resolution and Compatibility with Chain-of-Thought Models*: The fMRI signals used are slow hemodynamic responses, while chain-of-thought reasoning is a dynamic, sequential process. Aligning these integrated fMRI signals with an averaged representation of the thinking tokens presents a potential neuro-temporal mismatch. We recommend explicitly acknowledging this limitation in the Discussion section. It would be valuable to discuss future directions that could address this, such as utilizing neuroimaging techniques with higher temporal resolution (e.g., MEG or EEG) to capture the fast dynamics of reasoning processes, potentially enabling more fine-grained alignment with intermediate reasoning steps.*

Response:

Thank you for this important point. We agree that task-fMRI reflects a temporally integrated hemodynamic response, whereas chain-of-though reasoning may unfold as a fast, token-by-token dynamic process. Our current implementation therefore aligns fMRI signals with an aggregated representation (averaged thinking-token representations), while future works can investigate modalities with higher temporal resolution such as MEG or EEG to better track intermediate reasoning steps and potentially enable more fine-grained guidance with multi-step trajectories.

Revision

We have added the following discussion in the Discussion section:

"As fMRI signals are slow hemodynamic responses that may not reflect fast dynamical processes of chain-of-thought reasoning, we only consider aggregated representations, while future investigation on modalities with higher temporal resolution such as MEG or EEG is possible to overcome this limitation by tracking fast intermediate steps for potentially more fine-grained guidance."

11. I did not find the source code in the supplementary.

Response:

Thank you for pointing out this. In the previous submission, the source code was provided as a zip file in the supplementary materials, which may have issues for accessing. In this revision, we provide the link to the source code: https://github.com/pkuxmq/Brain-guided_LLM, which has also been included in the code availability statement.

2.3 To Reviewer 2

We sincerely thank the reviewer for the thoughtful and constructive evaluation of our work, and for recognizing the conceptual contribution, broad evaluation, and careful experimental design. In this revision, we have substantially strengthened the manuscript by adding new experiments, analyses, and explanations to address all concerns. Below we provide a detailed point-by-point response.

W1. While the authors frame this as a "paradigm shift" toward cognitively principled AI, the long-term scalability of the approach is questionable. *Given the high cost and logistical difficulty of acquiring high-quality fMRI data, it is unclear if this method can truly compete with large-scale behavioural fine-tuning or synthetic data generation. The authors should clarify the "neural data-efficiency" of their model – specifically, how much brain data is required to achieve gains that cannot be replicated through standard machine learning techniques.*

Response:

Thank you for this important point. We agree that the long-term scalability, comparison with standard machine learning techniques, and neural data efficiency deserve clearer discussion discussions.

First, we would like to clarify that the proposed approach does not aim to replace standard machine learning techniques; instead, we suggest neural signals as an additional and complementary source of supervision that can provide information not directly contained in behavioural languages. This is already reflected in our experiments of NARF+Label (Fig. 6), where neural guidance yields additional generalization gains beyond language supervision.

Moreover, to further evaluate the synergy of NARF and other techniques such as synthetic data, here we add a new experiment combining

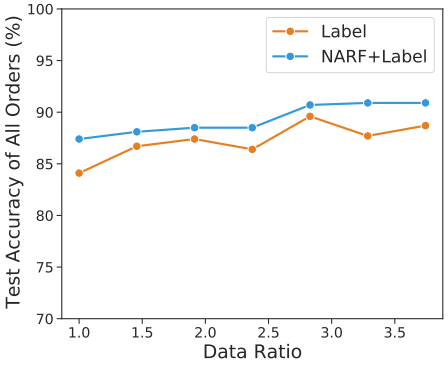

Fig. S12 Results of Combining NARF with Synthetic Data. Data ratio refers to the ratio of training data (problems from the fMRI dataset plus new synthetic data) over original problems from the fMRI dataset.

NARF+Label with additional synthetic training data (implementation details in Supplementary Information Section S11.6, quoted below). In this setting, only the original fMRI-derived problems contribute neural guidance, whereas the added synthetic data provide language supervision only. Fig. S12 presents the results of performance under different data ratio (total problems divided by instances from the fMRI dataset), showing that neural guidance continues to provide gains over language-only supervision across data ratios, even with limited fMRI data. The results suggest that the benefit of neural signals is orthogonal to synthetic data, and the proposed method aims to collaborate and complement rather than compete with large-scale behavioural fine-tuning or synthetic data generation.

With regard to the "neural data efficiency", our results show that even limited data can provide gains. In our current study, the neural signals are derived from only 70 deductive reasoning problems, and for each model only a small subset of those problems (around 3-20 problems that answered incorrectly by the model while correctly by human) directly contributes to guidance. Despite this limited scale, experiments still show measurable gains from neural guidance, demonstrating certain "data efficiency". While this is an initial empirical

indication rather than a full scaling law, as limited task-specific fMRI data prevents us from investigating a larger scale of data, we view the current results as evidence that limited neural data can be useful.

Considering the long-term scalability, while we acknowledge the difficulty of acquiring high-quality fMRI data, we expect our study can inspire and encourage the investigation and release of more task-specific brain data. Given that even limited task-fMRI data is possible to provide orthogonal improvement over behaviours for the task, we may expect the potential for scaling with more cognitive tasks as richer diverse task-specific neural datasets become available in the future, and we leave full characterization of neural-data scaling to future work.

Revision

We have added the new results and discussions in Supplementary Information Section S11.7 and Fig. S12, and clarified this complementary role more explicitly in Introduction and Discussion. We quote the revision below:

In Introduction:

"Taken together, our findings indicate that LLMs and the human brain share partially overlapping mechanisms for deductive reasoning, evident both in correlational representational alignment and in the functional gains achieved when neural signals are used to guide models. This work introduces a brain-signal-guided pathway to advance LLMs by integrating the structure of human cognitive neural data, complementary to behavioural supervision, marking a significant step towards cognitively principled and capable AI systems."

In Discussion:

"Our work provides an alternative perspective on learning intermediate representations directly from human neural signals rather than solely through language behaviour, offering a different possibility for "process supervision".

This approach doesn't compete with, but rather complements language behaviour supervision (Fig. 6; Sec. S11.7)."

In Supplementary Information:

"**S11.7 Combination with synthetic data**

We suggest neural signals as an additional source of information for representations that is complementary to common behavioural supervision, and the proposed method aims to collaborate and complement rather than compete with common machine learning techniques. To further evaluate the potential synergy of NARF and other techniques, we evaluate combining NARF+Label with additional synthetic data. Specifically, we generate synthetic data similar to the generation of test problems, while using different pseudowords and adjectives. The list of adjectives for syllogistic reasoning is: *ancient, modern, fierce, gentle, elegant, clumsy, swift, lazy, curious, mysterious, fragile, sturdy, noisy, silent, graceful, awkward, clever, foolish, generous, selfish*, and the list of comparative adjectives for transitive reasoning is: *easier, busier, noisier, safer, riskier, wiser, cleverer, luckier, healthier, cheaper, costlier, earlier, later, rarer, simpler, stricter, looser, fairer, nearer, further*. We also generate the same number of questions for all groups for a balanced data distribution. As the original fMRI dataset contains 70 problems, we generate 32, 64, 96, 128, 160, and 192 questions by gradually adding new instances to study the performance under different training data ratio over the original data. These questions only have language labels for supervision. We use the same training setting to train Mistral-7B for Label and NARF+Label under the same random seed, keeping the gradient accumulation step the same as 70.

Fig. S12 presents the results of performance over data ratio, showing that neural guidance continues to provide gains over language-only supervision

across data ratios, even with limited fMRI data. The results suggest that neural signals can provide orthogonal gains over common methods, with certain "neural data efficiency"."

W2. The study's empirical breadth is limited by the use of a single fMRI dataset *for both alignment and validation. While task-specific fMRI data for reasoning is scarce, the reliance on a singular source raises concerns regarding domain-generality. To establish NARI/NARF as robust NeuroAI tools, the authors should demonstrate their efficacy in other cognitive domains (e.g., language processing or speech), where larger, diverse neural datasets are more readily available. This would clarify whether the method captures a general principle of brain-AI integration or is specifically tuned to the unique structures of formal logic.*

Response:

Thank you for this valuable suggestion. We agree that validation on other fMRI datasets is important to demonstrate that our method is not specific to one experimental design.

To address this concern as directly as possible within the scope of the present study, here we add new experiments on an external dataset, HCP Relational Processing Task from the Human Connectome Project [3], as shown in Fig. S14. The HCP Relational Processing Task probes relational processing by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top images differ in (Fig. S14b). We adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a). Such different modalities extend our experimental settings to evaluate another abstract relational processing

ability with different stimulus modalities. This dataset differs from our original deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

We conduct both intervention and fine-tuning experiments on this task and dataset using five models from different families (Phi4-mini, Qwen2-7B, Llama3-8B, Gemma2-9B, Mistral-7B), and the implementation details are presented in Supplementary Information Section S13 (quoted below). As shown in Fig. S14c, human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges. While the overall intervention success rate on this dataset (around 80%) is lower than the original deductive reasoning setting (100%), it is likely due to an increased difficulty introduced by the modality shift, which reflects a boundary condition while remaining effective. In addition, Fig. S14d further shows that the general direction is possible to transfer to new, out-of-set problems. Finally, in a similar fine-tuning setting with evaluation on new questions under different orders (Fig. S14e), NARF complements language supervision with consistent gains over language supervision alone (Fig. S14f).

These results on an independent dataset show that the pipeline is not limited to one specific experimental design. Moreover, they show that our framework can extend to another reasoning-related task with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data.

As our present study focuses on the reasoning domain that goes beyond pure language processing, and our core claim in the paper is for reasoning, we mainly evaluate the related task rather than language processing or speech.

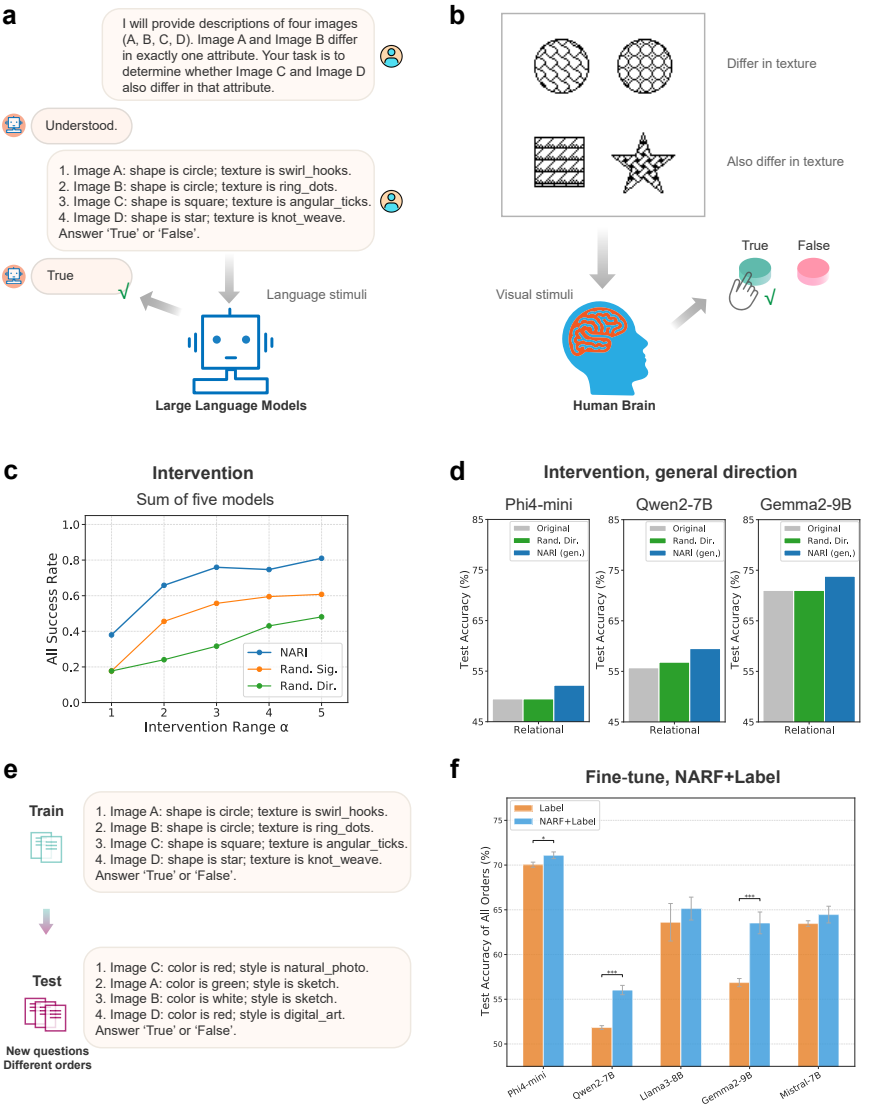

Fig. S14 Results on the Relational Processing Task from the HCP Dataset. (a-b) Illustration of the task for LLMs and humans. The task probes relational processing by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top image differ in. LLMs process problems through chat-formatted prompts with language-based descriptions of image attributes, generating direct linguistic responses. Human participants view two pairs of images (top and bottom), responding via button press. (c) Summary results of five LLM models for intervention on their incorrect problems. NARI achieves much higher success rates than baselines of random signals and the max-integration of random directions. (d) Results of the general direction on the new test data. (e) Illustration of the training and testing settings. (f) Testing accuracies of various models for language-label-only and NARF+Label fine-tuning.

46 Response

We agree that evaluating the method on other cognitive domains is important future direction. Our new results indicate that the method is possible to be extended to other tasks, suggesting the potential for investigation of more cognitive tasks in future work.

Revision

We have added these new results, discussions, and details in Supplementary Information Section S13 and Fig. S14, and briefly introduced this extension in the main text (Section 2.2). We quote the revision below.

The revision in Section 2.2 is:

"Extension to other neural datasets We further validate our method on another independent dataset, HCP Relational Processing Task from the Human Connectome Project [48] (Sec. S13, Fig. S14). This task probes relational reasoning by comparing relations from two pairs of images, and we adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a-b). This dataset differs from the deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

The results show that human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges (Fig. S14c), and the general direction of NARI is possible to transfer to new, out-of-set problems (Fig. S14d). At the same time, NARF complements language supervision with consistent gains (Fig. S14f). Full results and details can be found in Sec. S13.

The results show that our pipeline is not limited to one specific experimental design. Meanwhile, they show that our method can extend to other related

reasoning tasks with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data."

The revision in Supplementary Information Section S13 is:

"**S13 Extension to the HCP relational processing dataset**

In this section, to further validate our method on other independent datasets, we extend our experiments to HCP Relational Processing Task from the Human Connectome Project [3] (Fig. S14). The HCP Relational Processing Task probes relational reasoning by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top images differ in (Fig. S14b). We adapt it to large language models by presenting language descriptions of the attributes of the images to the models (Fig. S14a). Such different modalities extend our experimental settings to evaluate a different abstract relational processing ability with different stimulus modalities. This dataset differs from our original deductive reasoning dataset in both task and stimulus modality, thereby providing a meaningful external validation of the overall pipeline.

Details of the fMRI dataset and preprocessing The dataset is derived from the relational processing task in the Human Connectome Project (publicly available on www.humanconnectome.org), which originally contains more than 1,000 human subjects. We randomly select 20 subjects (IDs: 100610, 103111, 139637, 149236, 162733, 169444, 200008, 200513, 211619, 308129, 421226, 441939, 529549, 567759, 611938, 727553, 771354, 793465, 872158, 984472), whose behavioural accuracy ranges from 70% to 100% for this task. The dataset probes higher-order relational reasoning using visual stimuli that vary in shape and texture. It includes relational and matching (the control condition) blocks across two runs for each subject. For the relational block, each trial presents two pairs of images, where the top images differ in exactly

one attribute (shape or texture). Participants are required to judge whether the bottom images also differ in that attribute and respond via button press. Each subject has a total of 24 relational trials (2 runs, 3 relational blocks each run, and 4 trials each block).

While the dataset follows a block-designed task-fMRI paradigm, we extract the stimuli for each trial from the accompanying description file of experimental designs for each subject (which describes the file name of each stimuli) and the publicly available stimuli files (from the HCP_TFMRI_scripts). As the stimuli files are images, we manually annotate the shape and texture attributes of each image and convert them into language descriptions. There are 96 stimuli files in the scripts, while we find that only 47 stimuli files appear in fMRI experiments of the 20 subjects. There are total six possible shapes and six possible textures for images.

We use the fMRI data preprocessed by the Human Connectome Project, and estimate the response to each trial with GLMSingle. We use the average response time as the response time in GLMSingle. For the design matrix, we consider up to 3 conditions (used for cross-validation in GLMSingle): correct answer for questions with "yes" answer, correct answer for questions with "no" answer, and wrong answer, and extract single-trial responses for each problem with the design matrix.

We extract voxels from the estimated response based on locations of MD regions and the subject-specific functional localization process by contrasting relational blocks over matching blocks. We first use SPM12 to get the t-values of all voxels considering the contrast, and then select 10% of the most responsive voxels within the MD mask. These voxels are considered as vectors of brain representations for each human subject.

Details of model prompting We convert image stimuli to language descriptions for LLMs to process. Specifically, we convert the six possible shapes into: *circle, cross, triangle, square, star, hexagon*, and convert the six possible textures into: *ring_dots, angular_ticks, zigzag_dashes, solid_blocks, knot_weave, swirl_hooks*. The task prompt for the model is:

"*You are an expert in relational reasoning. I will provide descriptions of four images (A, B, C, D), which describe two attributes of each image: shape and texture. Image A and Image B differ in exactly one attribute (either shape or texture). Your task is to determine whether Image C and Image D also differ in that attribute. For example, if Image A and Image B have different shapes, then you should determine if Image C and Image D also have different shapes. Or if Image A and Image B have different textures, then you should determine if Image C and Image D also have different textures. If the statement is true, you should answer 'True'; otherwise answer 'False'. Only answer 'True' or 'False' at the beginning of the response.*"

After the model's round saying "*Understood. I will analyze the attributes and perform relational reasoning to judge the statement. I will only answer 'True' or 'False' at the beginning of the response.*", we present the descriptions as follows:

"*1. Image A: shape is circle, texture is solid_blocks.\n 2. Image B: shape is circle, texture is ring_dots.\n 3. Image C: shape is triangle, texture is zigzag_dashes.\n 4. Image D: shape is hexagon, texture is zigzag_dashes.\n For the attribute that Image A and Image B differ in (either shape or texture), do Image C and Image D also differ in that attribute? Only answer 'True' or 'False' at the beginning of your response.*"

Then we extract the first token of the model's response.

Generated test dataset For testing of LLMs, we generate new questions with different descriptions of attribute types and attributes. Specifically, we replace "shape" and "texture" by two random attribute type in *[color, quality, geometry, style]*, and each of them has the following possible attributes: *color: [red, green, blue, yellow, black, white]; quality: [normal, noisy, blurred, underexposed, overexposed, compressed]; geometry: [rotated, translated, zoomed_in, zoomed_out, flipped, sheared; style: [sketch, watercolor, oil_painting, digital_art, cartoon, natural_photo]*. There are four conditions: true/false $\times$ first/second attribute is different. For the test data, we generate 100 problems for each condition, leading to 400 problems; for the validation data for fine-tuning, we generate 20 problems for each condition, leading to 80 problems. When evaluating fine-tuned models, we also enumerate all possible permutations for the order of descriptions of four images (Fig. S14e), with finally $400 \times 24 = 9600$ test data.

Implementation details of NARI and NARF We leverage the similar experimental settings as the deductive reasoning task. For NARI, the maximum gradient descent steps are set as 50. For the general direction of NARI, we take the average of successful intervention directions with $\alpha = 3$ (for Phi4-mini, we take $\alpha = 5$), normalize the direction for each layer to the standard deviation scale, and the scale γ is searched between 0.5 and 10 with the interval 0.5. For the combination of NARF and language labels, we use all 96 questions from the HCP stimuli files with only part of them having fMRI signals. We select 10 subjects for fine-tuning based on the intervention success rate, and only perform guidance for questions that the model can be intervened to the correct answer. We start training of NARF+Label from the language-label-trained models, except for Phi4-mini that we find starting from the original model has better performance. λ_{narf} is set as 0.1 for Phi4-mini

and Llama3-8B, 0.05 for Qwen2-7B, and 0.01 for Mistral-7B and Gemma2-9B. We also run each experiment for 5 times with the same random seeds: 0, 1, 42, 1000, 2024, and report statistical significance based on two-sided paired t-test among the 5 runs.

Results We conduct intervention and fine-tuning experiments on this task and dataset using five models from different families (Phi4-mini, Qwen2-7B, Llama3-8B, Gemma2-9B, Mistral-7B). As shown in Fig. S14c, human neural activations from this dataset can also induce effective representation intervention directions, significantly outperforming both baselines (model representational structure alone / random directions) across perturbation ranges. While the overall intervention success rate on this dataset (around 80%) is lower than the original deductive reasoning setting (100%), it is likely due to an increased difficulty introduced by the modality shift, which reflects a boundary condition while remaining effective. In addition, Fig. S14d further shows that the general direction is possible to transfer to new, out-of-set problems. Finally, in a similar fine-tuning setting with evaluation on new questions under different orders (Fig. S14e), NARF complements language supervision with consistent gains over language supervision alone (Fig. S14f).

These results on an independent dataset show that our pipeline is not limited to one specific experimental design. Meanwhile, they show that our method can extend to other tasks with different stimulus modalities, suggesting a promising way to incorporate more potential neural signal data."

W3. The functional utility of NARI is partially obscured by the results in Figures 4h and 4k. *For certain models, the performance gains over "Random Direction" baselines appear marginal. If a stochastic perturbation of the same magnitude performs comparably to a brain-derived direction,*

it suggests that NARI might be acting as a generic form of representation-space regularization rather than a specific cognitive guide. Further clarification is needed on how test accuracy scales as a function of the intervention parameter (α) for these random directions. The alignment of the brain-trained models actually also decreases in some cases (Fig. S8).

Response:

Thank you for this important point. First, following your suggestion, we supplement how test accuracy changes as a function of the intervention parameter α for both brain-derived directions and random directions in Fig. S6. As shown in the results, random directions produce little or no change in performance across α, while brain-derived direction mostly has significant influence (typically, performance improves at moderate scales and then declines due to over intervention when α becomes too large). This suggests that brain-derived directions acts in a more reasoning-relevant subspace, which differs from random directions, and NARI is not merely a regularization similar to stochastic perturbation.

Second, we agree that the gains in Fig. 4h and 4k are not uniform across all models. We would like to clarify that this variability is consistent with the literature in representation steering, whose effectiveness has been shown to depend on the model and is not uniformly reliable [6, 7]. We therefore interpret the phenomenon that certain models do not have good performance as reflecting a model-dependent limitation of steering methods, rather than evidence that the brain-derived direction behaves like stochastic perturbation. This is possibly a shared limitation of steering methods and may be improved by more advanced steering methods in the future work. While in this work, we mainly aim to verify the plausibility of brain guidance through both intervention and fine-tuning experiments.

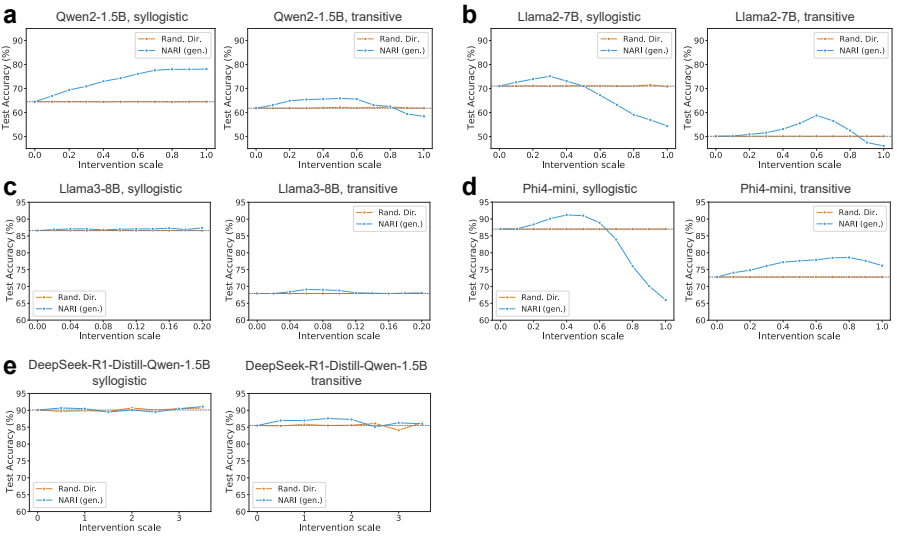

Fig. S6 Analysis Results of General Direction Intervention Experiments. Test accuracy of NARI (gen.) and random direction with different scales for various models.

Finally, for the alignment score of brain-trained models, we would like to clarify that more accurate model does not necessarily have higher global alignment with human. This is because 1) humans are not perfect (no subject achieves 100% accuracy; the average of human accuracy in the dataset is 86.4%) and 2) functional guidance and global alignment metrics are not necessarily monotonically coupled. Our method applies neural guidance primarily to model's incorrect problems, whereas standard alignment score compute Pearson correlation over all samples. As a result, this score does not necessarily increase, while the sample-wise cosine similarity (our objective) increases. It is reasonable that some alignment score decreases, which does not influence the effectiveness of brain-derived directions.

Revision

We have added the analysis results and discussions of NARI (gen.) in Supplementary Information Section S10.3 and Fig. S6. We have also extended the

discussion of alignment scores in Supplementary Information Section S11.3. We quote the revision below:

"**S10.3 Analysis results of NARI (gen.) over intervention scale**

In this section, we further provide analysis results of how test accuracy of NARI (gen.) scales as a function of the intervention scale α. As shown in Fig. S6, random directions produce little or no change in performance across α, while brain-derived direction mostly has significant influence (typically, performance improves at moderate scales and then declines due to over intervention when α becomes too large). This suggests that brain-derived directions acts in a more reasoning-relevant subspace, which differs from random directions.

We also notice that the applicability of representation steering varies for different models. Actually, it has been shown that the effectiveness of representation steering depends on the model and is not uniformly reliable [6, 7]. It may be improved by more advanced steering methods in the future work."

And the revision in Supplementary Information Section S11.3 is:

"Indeed, more accurate model does not necessarily have higher global alignment with human, because 1) humans are not perfect (no subject achieves 100% accuracy; the average of human accuracy in the dataset is 86.4%) and 2) functional guidance and global alignment metrics are not necessarily monotonically coupled. Our method applies neural guidance primarily to model's incorrect problems, whereas standard alignment score compute Pearson correlation over all samples. As a result, this score does not necessarily increase, while the sample-wise cosine similarity increases."

W4. To quantify neural predictivity*, the authors take the maximum alignment across all evaluated layers as the final "brain-score," following Schrimpf et al. (2021). However, this approach has been iterated on since to resolve issues regarding the "oracle" nature of selecting the best layer on the test set itself*

which might inflate predictivity scores. Contemporary studies in NeuroAI mitigate this selection bias by either i) utilizing functional localizers [1] to isolate model units that are specifically selective for a given functionality (e.g., deductive reasoning), mirroring how functional regions are identified in the human brain; or ii) using a validation or "layer-selection" set to commit to a specific layer before evaluating performance on a held-out test set. These approaches ensure that the final evaluation is independent of the layer or unit selection process, thereby providing a more scientifically valid estimate of representational alignment. In our experience, with a sufficiently large dataset these results will be consistent but since this is a new domain of NeuroAI inquiry, we feel it is worthwhile to double-check.

Response:

Thank you for this valuable suggestion. We agree that using the maximum alignment across layers can introduce potential selection bias, and a more scientifically valid estimate is important.

Following your suggestion, we have re-calculated neural predictivity using a functional localization procedure following recent NeuroAI practice [8, 9] (details in Methods and Supplementary Information Section S3.1, quoted below), and updated the corresponding results in Fig. 2 as well as Supplementary Information Fig. S2 (we still report layerwise results in Supplementary Information). The overall pattern and main conclusions remain unchanged. Specifically, while the absolute alignment scores are slightly lowered (e.g., for all reasoning questions, from 0.9 (average of previously six models) to 0.76 (average of updated ten models)), all previous results still hold: (1) across all trials, predictivity is significantly higher in reasoning and MD regions than in language regions, (2) analyzing reasoning types separately, ceiling-normalized predictivity drops for each type (to around 0.27), (3) predictivities of trained

models are significantly higher than untrained models, (4) each type remains regionally specific (deductive reasoning $\geq$ language and MD). We believe that this updated analysis further strengthens the validity our conclusions.

Revision

We have updated neural predictivity results in Fig. 2 and Supplementary Information Fig. S2 based on the latest functional localization procedure, and also updated related descriptions in Methods and Supplementary Information Section S3.1. We quote the revision below.

In Methods:

"Neural predictivity is calculated based on functionally localized model units following the latest practice [38, 39]. The calculation process involves the following steps. First, we localize model units responsive to deductive reasoning by constrasting activations between the reasoning condition and a control condition (only reading premises; Sec. S3.1). We choose top-N most responsive units from both attention outputs and layer outputs [23] at the last token, where N corresponds to the width of the model's layer. The activations from these units for each problem are extracted as model representations. In addition to this, we also investigate representations from each layer."

In Supplementary Information Section S3.1:

"**S3.1 Prompt for the control condition**

To functionally localize model units responsive to deductive reasoning for neural predictivity calculation, we further design a control condition to only read premises. Under this condition, the task prompt of syllogistic premises to the model is:

"*I will present three premises that describe a series of relationships among monosyllabic pseudowords and adjectives. You should just read the descriptions and do nothing. Only respond 'I have read the premises' after my messages.*"

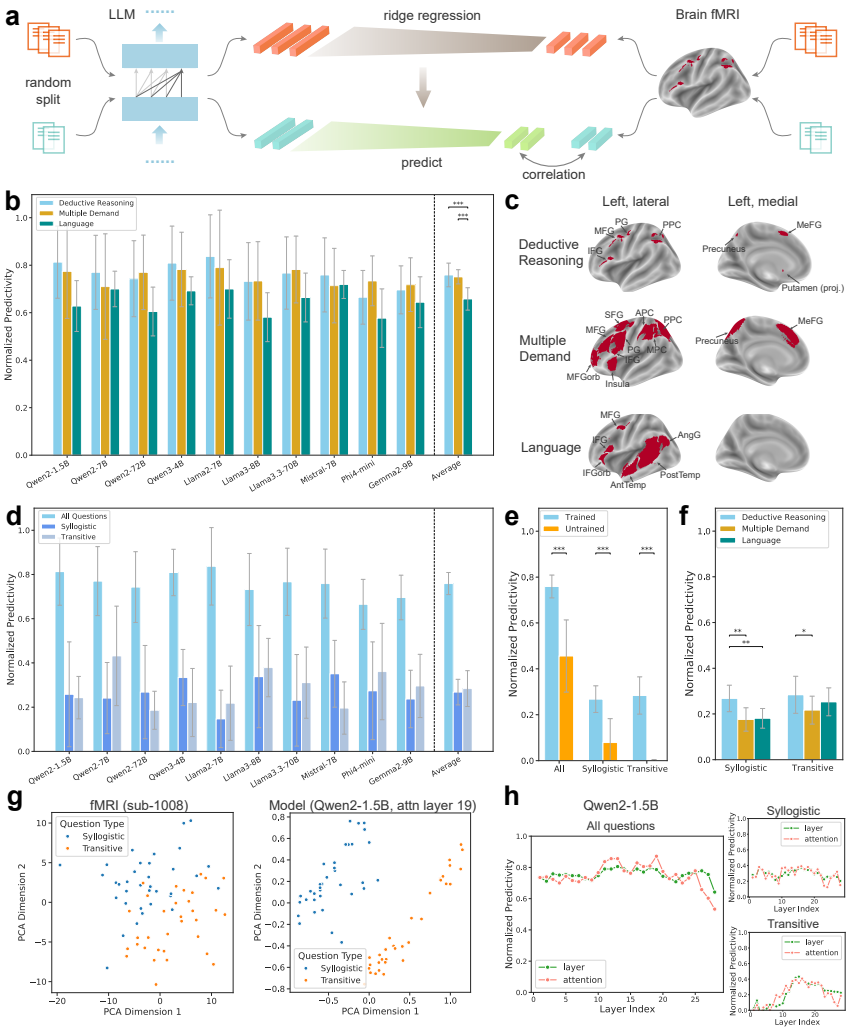

Fig. 2 Results of Neural Predictivity. (a) Neural predictivity calculation pipeline: Fitting ridge regression to map model representations to brain activations, and computing held-out correlation scores divided by ceiling values (median across voxels and participants). (b) Neural predictivity scores of different models across brain regions (deductive reasoning, multiple-demand (MD), language) for all reasoning problems (both syllogistic and transitive), showing significantly higher alignment in deductive reasoning and MD regions (error bars: inter-subject variability for each model's results and inter-model variability for the Average results). (c) Anatomical locations of deductive reasoning, MD, and language regions. (d) Type-specific predictivity scores (syllogistic or transitive). (e) Comparison between trained and untrained (weights randomly initialized) models, showing significantly higher predictivity of trained models. (f) Regional predictivity scores stratified by reasoning type. (g) Representational analysis revealing type separation among model and brain representations. This separation largely accounts for the high predictivity under all questions. (h) Layer-wise predictivity scores under various settings.

For transitive premises, the description is also slightly different as above. After the model's round saying "*Understood.*", only the three premises are presented."

W5. The remarkably high neural predictivity observed in untrained models *(Figure 2e) is a potential "red flag" that may point to methodological artifacts. Our intuition is that randomly initialized weights should not, in principle, align closely with complex human reasoning. This might suggest train-test contamination within the k-fold cross-validation, potentially caused by shuffling samples with strong temporal or contextual dependencies. If samples from the same stimulus block appear in both sets, the regression may be fitting to low-level signal characteristics or scanner noise rather than genuine neural representations.*

Response:

Thank you for this important point. We agree that high neural predictivity in untrained models requires more detailed explanation.

First, we note that the predictivity of untrained models is largely reduced in the updated neural predictivity results calculated based on functional localization (shown in Fig. 2). In particular, when syllogistic and transitive reasoning are analyzed separately, the neural predictivity of untrained models becomes near zero, whereas trained models remain substantially higher. This pattern makes it unlikely that the results are driven by low-level scanner noise or a generic artifact in the regression pipeline.

Second, for the setting of all reasoning questions, the above-zero predictivity for untrained models is likely driven by the separation between reasoning types. As shown in Fig. 2g, brain representations exhibit certain type separation. At the same time, the untrained models in our analysis retain the trained tokenizer (our untrained model refers to

untrained weights for the main model), so coarse linguistic differences between descriptions of syllogistic and transitive questions may still induce a corresponding separation in model representations. This potential confounder leads to above-zero predictivity score of untrained models. Despite this, trained models have much larger scores than untrained models, indicating that models capture brain representations beyond these confounding and noisy causes.

Revision

We have added this discussion of results of untrained models in Supplementary Information Section S7. We quote the revision below:

Discussion on results of untrained models We note that while untrained models have near zero neural predictivity for each type of reasoning, they have above-zero predictivity scores under the setting of all reasoning questions (Fig. 2e). This is likely driven by the separation between reasoning types. As shown in Fig. 2g, brain representations exihibit certain type separation. At the same time, the untrained models in our analysis retain the trained tokenizer, so coarse linguistic differences between descriptions of syllogistic and transitive questions may still induce a corresponding separation in model representations. This potential confounder leads to above-zero predictivity score of untrained models. Despite this, trained models have much larger scores than untrained models, indicating that models capture brain representations beyond these confounding and noisy causes.

W6. The PCA and KDE visualizations (e.g., Figure 5c) primarily illustrate the consequences of fine-tuning (i.e., better class separation) rather than the mechanisms of neural guidance. *These plots confirm that the model's accuracy has improved, but they do not sufficiently explain how the neural signals steered the model toward a more robust logical manifold.*

Response:

Thank you for this important question. We agree that the PCA and KDE visualization in Fig. 5c primarily illustrate the consequence of fine-tuning rather than the steering process.

To clarify this distinction, we note that the steering mechanism of our method is most directly reflected in the intervention setting, where the brain-guided direction is explicitly computed and applied in the model representation space. In particular, Fig. 4g and Supplementary Information Fig. S4 visualize the actual neural-guided steering direction and its effect for intervention in the model representation space. These figures provide the most direct illustration of how neural signals steer the model.

On the other hand, in fine-tuning experiments, the representational knowledge provided by neural guidance is internalized into model parameters rather than applied as an explit steering vector. For this reason, it is much harder to directly visualize. Therefore, Fig. 5 and Fig. 6 mainly show the resulting representational reorganization as the consequence of fine-tuning, showing that model parameters have learned to optimize the representation space with neural guidance.

P1. There is a lack of uniformity in the models presented across the primary figures. *Figure 4 focuses on Qwen2 and Llama2/3; Figure 5 adds Mistral; and Figure 6 introduces a wide array of Instruct-tuned models (Phi, Gemma, etc.). This shifting selection makes it difficult to assess the consistency and robustness of the results across the entire study.*

Response:

Thank you for this important point. We agree that a clearer and more consistent presentation is important for the entire study. In the revised manuscript, we have updated Fig. 2, Fig. 4, and Fig. 5 to include more models.

Specifically, Fig. 2 now presents neural predictivity results of all ten models considered in this work (as in Fig. 6). We also supplement the results of Phi-4-mini-Instruct to Fig. 4 and Fig. 5, as well as to the related supplementary figures (Figs. S5, S7, and S9), so that these figures consistently report results for six models. This six-model subset was chosen for an evaluation reason: they have incorrect answers for both syllogistic and transitive questions from the fMRI dataset. This enables better evaluation and analysis of steering for different reasoning types. The remaining four models, on the other hand, only have incorrect answers for transitive questions, and we include them in NARF+Label experiments to demonstrate the general applicability and generalization of neural guidance.

Revision

We have updated Fig. 2, Fig. 4, Fig. 5, and Supplementary Information Figures S5, S7, and S9. We have also updated the description of models in Results, and we quote the revision below.

In Section 2.1:

"In parallel, we evaluate ten open-source LLMs spanning 1.5B-72B parameters from different families: Qwen2-1.5B-Instruct, Qwen2-7B-Instruct, Qwen2-72B-Instruct [31], Qwen3-4B [32], Llama2-7B-Chat [33], Llama3-8B-Instruct, Llama3.3-70B-Instruct [34], Mistral-7B-Instruct [35], Phi-4-mini-Instruct [36], and Gemma2-9B-it [37]"

In Section 2.2.1:

"We consider six models (Qwen2-1.5B/7B-Instruct, Mistral-7B-Instruct, Llama2-7B-Chat, Llama3-8B-Instruct, Phi-4-mini-Instruct) that have errors in both syllogistic and transitive questions. The results of six models demonstrate that NARI achieves a 100% success rate ..."

P2. Figure S8 indicates that in some instances, the alignment of brain-trained models actually decreases. *This counterintuitive finding deserves a more thorough discussion in the main text, as it complicates the narrative that alignment and performance are linearly coupled.*

Response:

Thank you for this important point. We agree that the relation between reasoning performance and brain alignment deserves more discussions.

We clarify that more accurate model does not necessarily have higher global alignment with human, because 1) humans are not perfect (no subject achieves 100% accuracy; the average of human accuracy in the dataset is 86.4%) and 2) functional guidance and global alignment metrics are not necessarily monotonically coupled. First, our method applies neural guidance primarily to model's incorrect answers, inducing a task-relevant guidance on the subset of cases that require correction rather than maximizing a global correlation metric, whereas standard alignment score compute Pearson correlation over all samples. As a result, the global score does not necessarily increase, while the sample-wise cosine similarity (our objective) increases (Fig. S9a). Second, the decrease of the Pearson-correlation-based alignment is also observed in language-supervision-based fine-tuning (Fig. S9b), further supporting that, given imperfect humans, performance and global alignment metrics are not necessarily monotonically coupled.

Revision

We have added this extended discussion in Supplementary Information Section S11.3, with reference in the main text. We quote the revision below:

In Section 4.6 (Methods):

"This objective induces a task-relevant neural guidance on the subset of cases that require correction rather than maximizing a global correlation

metric, and therefore is not necessarily monotonically coupled with global alignment with humans but related to sample-wise similarity (Sec. S11.3)."

In Supplementary Information:

"Indeed, more accurate model does not necessarily have higher global alignment with human, because 1) humans are not perfect (no subject achieves 100% accuracy; the average of human accuracy in the dataset is 86.4%) and 2) functional guidance and global alignment metrics are not necessarily monotonically coupled. Our method applies neural guidance primarily to model's incorrect problems, whereas standard alignment score compute Pearson correlation over all samples. As a result, this score does not necessarily increase, while the sample-wise cosine similarity increases."

Q1. Was there a theoretical or empirical rationale for focusing the intervention exclusively on the outputs of the attention modules and layers*, rather than including the outputs of the Feed-Forward Networks (FFNs) for example?*

Response:

Thank you for this insightful question. Our previous choice to perform intervention on attention outputs was motivated primarily by the prior inference-time intervention practice [5, 10], as well as the role of attention in modeling token relationships that is critical for deductive reasoning. It is also possible to intervene in other components, such as the layer outputs or FFN outputs. In additional experiments, we test intervention on layer outputs and we find that we can similarly achieve 100% intervention success rate together with similar improvements for the general direction setting. This suggests that our method is not limited to the attention output. Future work can investigate more refined designs.

Revision

We have added this discussion in Supplementary Information Section S6. We quote the revision below:

"We primarily followed the prior practice of inference-time intervention [11, 12] to performance intervention on attention outputs. It is also possible to intervene in other components, such as the layer outputs or FFN outputs. We test intervention on layer output and we can similarly achieve 100% intervention success rate together with similar improvements for the general direction setting. This suggests that our method is not limited to the attention output."

Q2. The results in Figure S5d are really interesting if significant. *Could the authors provide a formal statistical test for the significance of this result and explain why such a notable finding was relegated to the Supplementary Information?*

Response:

Thank you for this important suggestion. Here we perform a formal statistical test for the updated subject-level analysis results (with inclusion of Phi4-mini). The resulting correlation and statistical significance is $r = 0.61, p = 0.059$ across the 10 subjects. Additionally, we find that this correlation is sensitive to model families, where Llama2-7B and Llama3-8B contribute most to the overall correlation ($r = 0.68$ for Llama2-7B, $r = 0.50$ for Llama3-8B, and $r = 0.09$ for the sum of remaining four models). Therefore, we consider these results are suggestive for the correlation, while not statistically significant or sufficiently robust. For this reason, we briefly mention the suggestive correlation in the main text while leaving the results in the Supplementary Information.

Revision

We have added this statistical result and discussion in Supplementary Information Section S10.2. We quote the revision below:

"Considering the sum of six models, the subject-specific success rate shows quite strong positive correlation with individual human accuracy across the 10 subjects ($r = 0.61$, $p = 0.059$, Fig. S5d), suggesting that higher-performing subjects can not only provide more guiding problems (they answered correctly) for intervention but also induce more effective neural-guided directions for model representations. While we also find that this correlation is sensitive to model families, where Llama2-7B and Llama3-8B contribute most to the overall correlation ($r = 0.68$ for Llama2-7B, $r = 0.50$ for Llama3-8B, and $r = 0.09$ for the sum of remaining four models). These results suggest a potential while sensitive correlation."

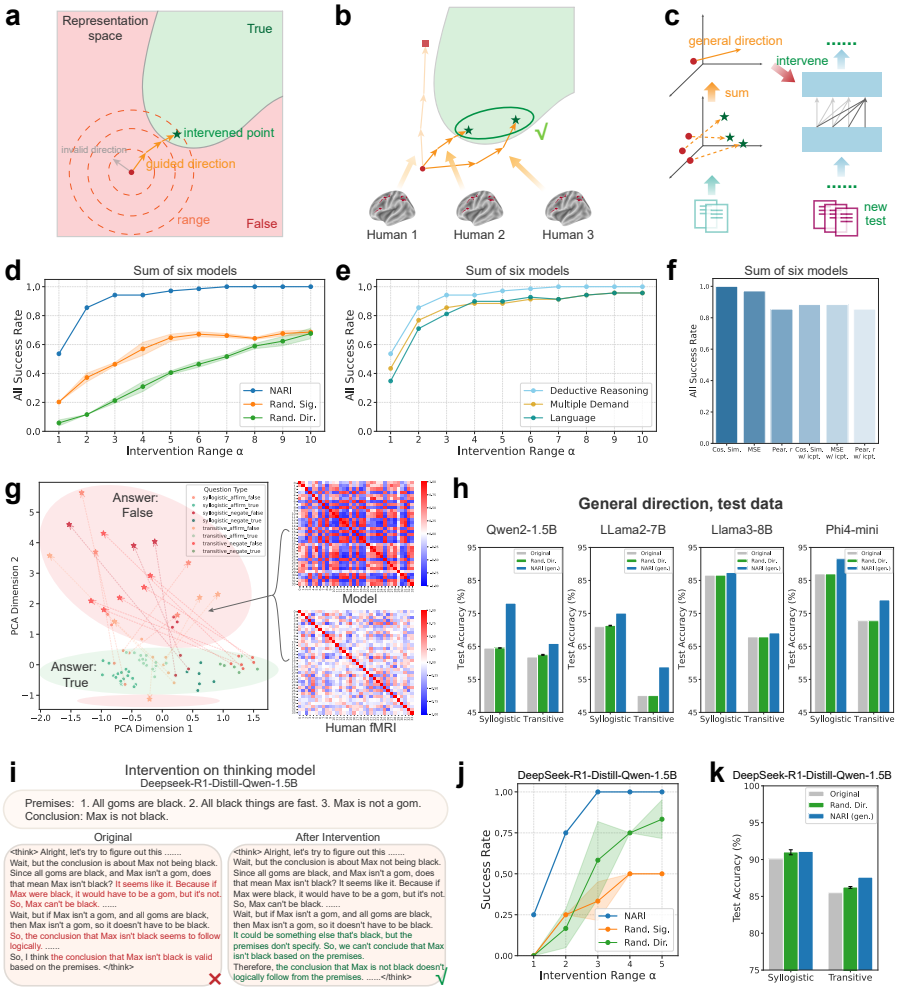

Fig. 4 Results of Representation Intervention Experiments. (a) Illustration of the representation intervention paradigm. Effective directions in the representation space can guide the representation of incorrect responses toward correct model outputs, with smaller required modification ranges indicating higher precision of the direction. (b) Different fMRI signals from various human subjects may result in diverse intervention results. We integrate all human subjects and consider a problem successful if a valid intervention exists. (c) We integrate successful directions into a general direction, enabling performance improvement on new test problems through inference-time intervention. (d) Summary results of six LLM models for intervention on their incorrect problems. Our Neural Activation guided Representation Intervention (NARI) achieves a 100% success rate and much higher rates in small range restrictions than baselines of random signals (eliminate the structure of fMRI representations and only relies on model representational structures) and the max-integration of random directions. (e) Comparison results of different brain regions. Deductive reasoning regions yield higher success rates than MD and language networks. (f) Comparison results of different objectives for direction calculation (*i.e.*, gradients), including cosine similarity (Cos. Sim.), mean square error (MSE), pearson r (Pear. r), and their conterpart with intercept (icpt.). (g) Visualization of the intervention process for Qwen2-1.5B-Instruct in the representation space (attention output of the 3/4-th layer). The joint effect of the structures of model and fMRI representations leads to successful intervention directions. (h) Results of the general direction on the new test data. (i) Demonstration of adapting NARI to chain-of-though reasoning models, enabling intervention on the language thinking process. (j) Intervention results on incorrect problems of the thinking model DeepSeek-R1-Distill-Qwen-1.5B. (k) General direction results on the thinking model DeepSeek-R1-Distill-Qwen-1.5B.

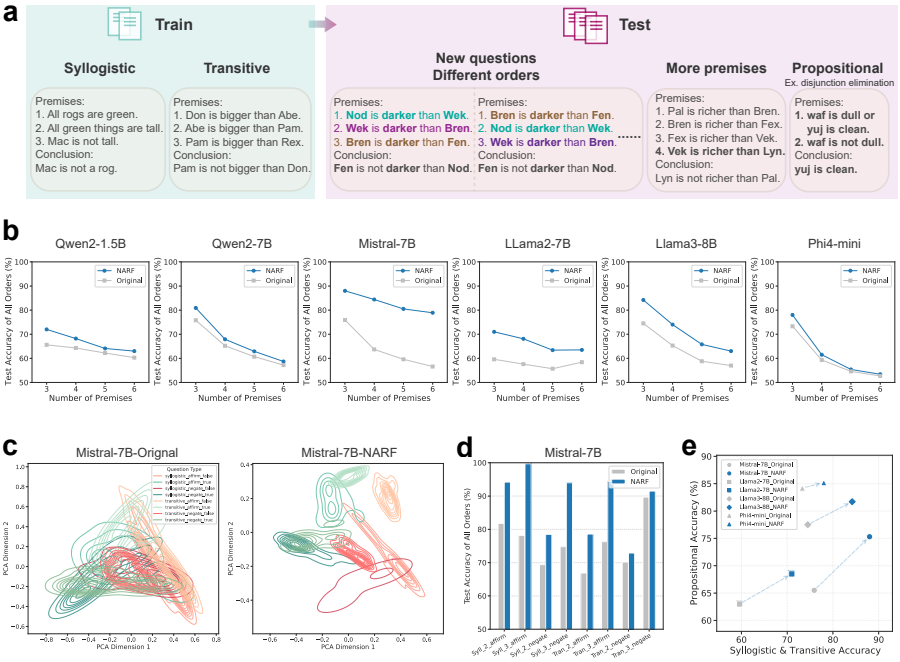

Fig. 5 Results of Parameter Fine-tuning Experiments. (a) Illustration of the training and testing settings. Training data includes syllogistic and transitive reasoning problems from the fMRI dataset, while we test models on new stimuli with varying premise orders, extended premise numbers, and different reasoning types. (b) Persistent testing accuracy improvement of various models on new data across different premise configurations, demonstrating robust generalization. (c) Kernel-density-estimation-based visualization of the model representations (attention output of the 3/4-th layer) on testing data reveals increased separation between problems with "True" and "False" labels in the representation space after fine-tuning. (d) Testing results on different subtypes of reasoning problems demonstrate consistent improvements (full results in Fig. S7). (e) Test results on propositional reasoning problems. Improvements of the model's behaviour generalize to other types of reasoning.

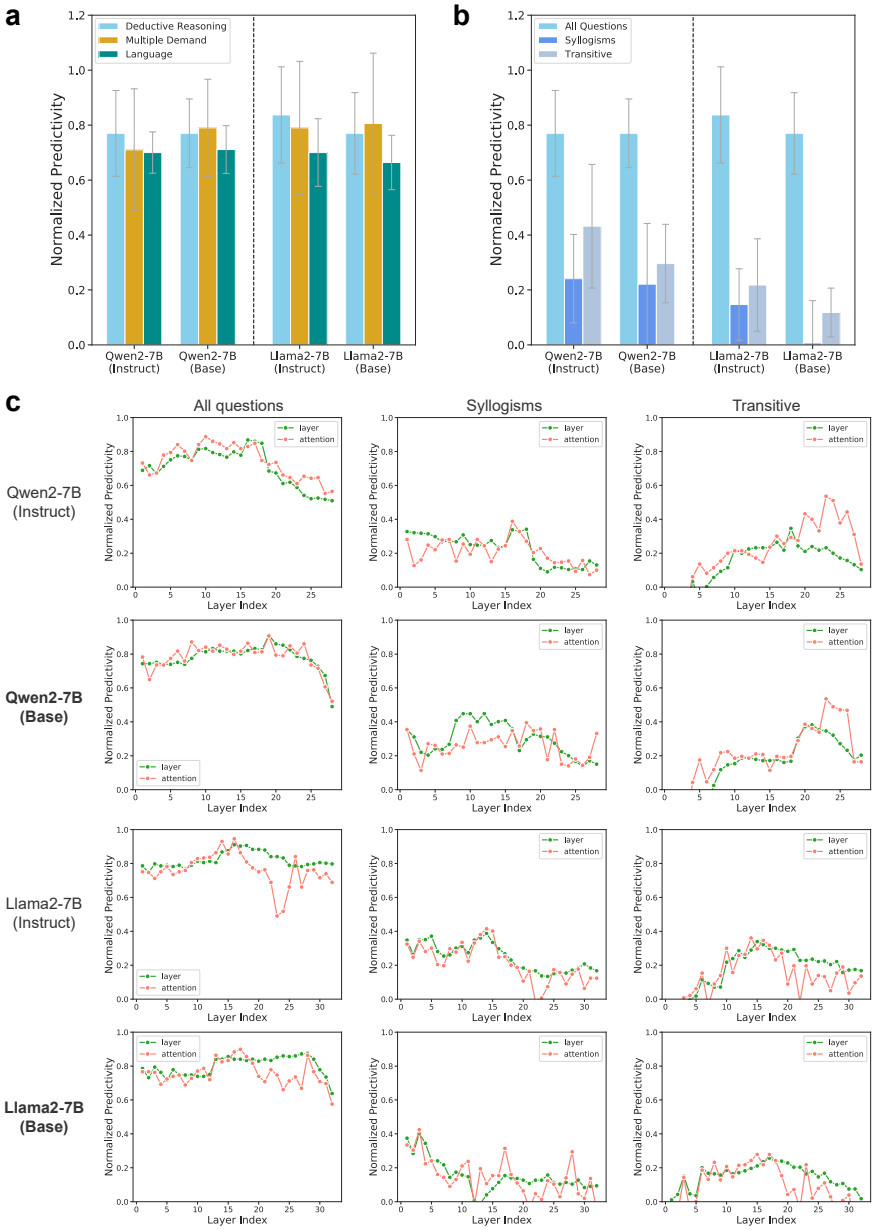

Fig. S2 Comparison Results of Neural Predictivity between the Instruction-tuned and Base Models. (a) Neural predictivity results for different brain regions considering all reasoning problems. (b) Neural predictivity results for different reasoning types. (c) Layer-wise neural predictivity results.

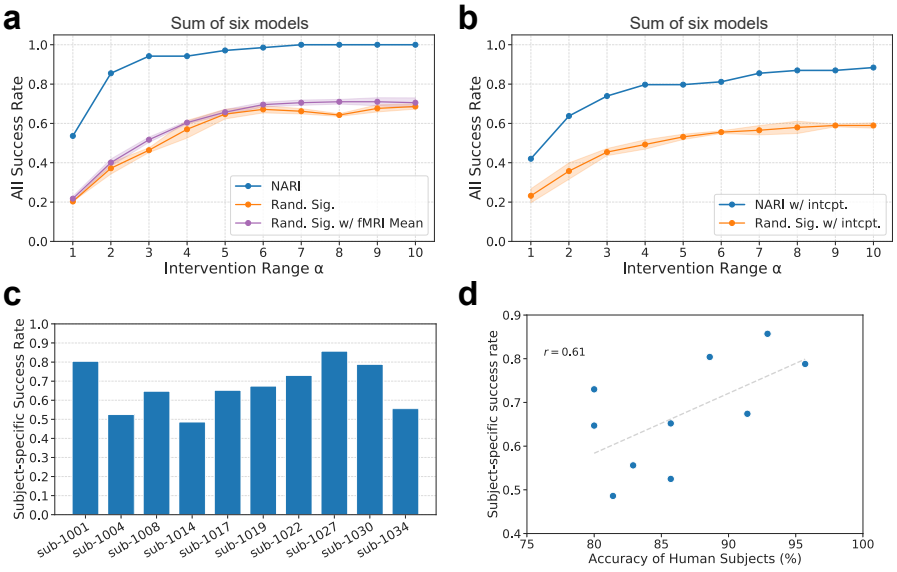

Fig. S5 More Analysis Results of Representation Intervention Experiments. (a) Results with the additional systematic shift term during direction calculation. "Rand. Sig. w/ fMRI Mean" refers to replacing fMRI signals with random signals whose mean vector keeps the same as fMRI signals. (b) Results without the additional systematic shift term, *i.e.*, calculating similarity metrics with the intercept term. (c) Subject-specific success rate of different human subjects. It is also the sum of five models. (d) The relation between subject-specific success rate and the accuracy of human subjects.

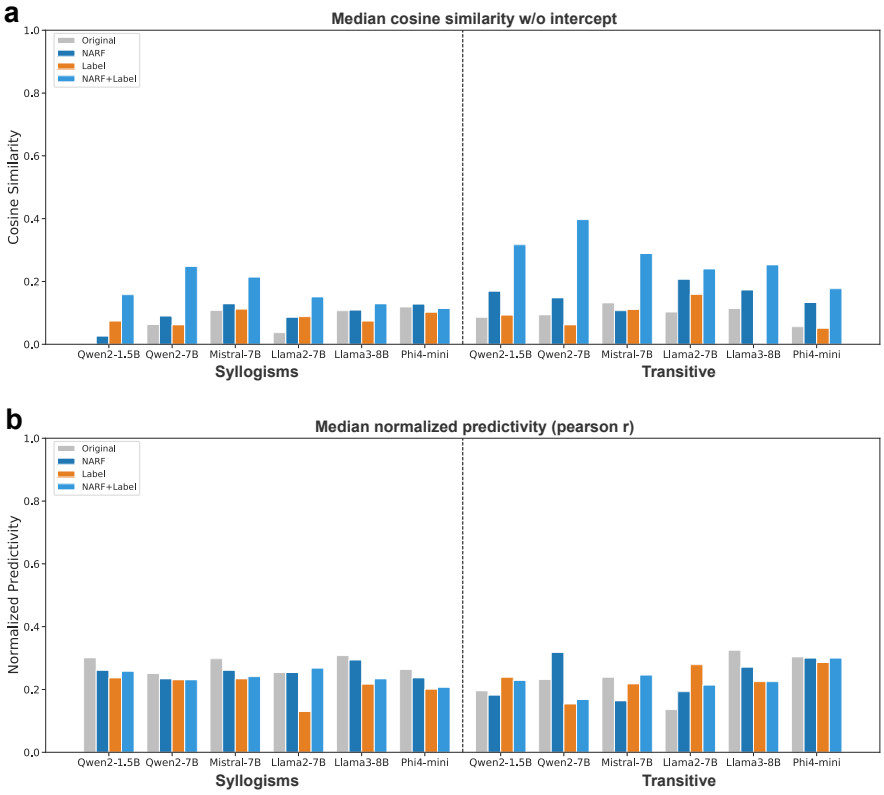

Fig. S9 Results of Similarity Metrics with Human Neural Activations after Fine-tuning. (a) Results of the intercept-excluded similarity metric (our objective), which removes the intercept term and is based on cosine similarity. (b) Results of the similarity metric following the predictivity calculation, which includes the intercept term and is based on Pearson r correlation among data samples.

References

[1] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

[2] 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 (PNAS)*, 118(45):e2105646118, 2021.

[3] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[4] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 2024.

[5] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[6] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[7] Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability

challenges in steering language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[8] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[9] 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 *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[10] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *International Conference on Machine Learning (ICML)*, 2024.

Response for "Beyond representational alignment with brain-guided language models for robust reasoning"

Mingqing Xiao[1,3], Kai Du[2*] and Zhouchen Lin[1*]

[1]State Key Lab of General AI, School of Intelligence Science and Technology, Peking University, Beijing, China.
[2]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China.
[3]Microsoft Research Asia.

*Corresponding author(s). E-mail(s): kai_du@tsinghua.edu.cn; zlin@pku.edu.cn;
Contributing authors: mingqing_xiao@pku.edu.cn;

We sincerely thank the Editor for handling our manuscript and the two reviewers for their careful evaluation, insightful comments, and constructive suggestions. In response to remaining reviewer comments, we have revised the manuscript and Supplementary Information to address all comments. Below, we provide a point-by-point response to each comment.

1 To Reviewer 1

We sincerely thank the reviewer for the thoughtful and constructive assessment of our work, and we are glad that our revision has resolved most points. We have further revised Supplementary Information to address the remaining small additions.

***1. An empirical sanity check**: apply instance-specific NARI to a random sample of initially-correct problems and report the flip-to-incorrect rate.*

Response:

Thank you for this valuable suggestion. To address this comment, we supplement an empirical sanity check for Qwen2-1.5B-Instruct. We apply a similar instance-specific NARI procedure to initially-correct problems with 50 iterations and different α, and report the average flip-to-incorrect rate (mean across subjects). The results show moderate flip-to-incorrect rates (1.6% for $\alpha = 1$, 13.0% for $\alpha = 2$, 23.6% for $\alpha = 3$, 28.9% for $\alpha = 4$, 32.0% for $\alpha = 5$, 33.5% for $\alpha = 6$, 35.1% for $\alpha = 7$, 36.6% for $\alpha = 8$, 37.0% for $\alpha = 9$, and 37.6% for $\alpha = 10$), which is consistent with our explanation of experimental logic that intervention on correct answers can in principle have both outcomes. The flip rate is dependent on the relative position of model and human representations compared to the decision boundary, as well as the robustness of the model. It cannot suggest meaningful conclusions.

Revision

We have added the results in Supplementary Information Section S12. We quote the revision below:

"For sanity check, we also apply a similar instance-specific NARI procedure to initially-correct problems with 50 iterations and different α, and report the average flip-to-incorrect rate (mean across subjects). The results show moderate flip-to-incorrect rates (1.6% for $\alpha = 1$, 13.0% for $\alpha = 2$, 23.6% for $\alpha = 3$, 28.9% for $\alpha = 4$, 32.0% for $\alpha = 5$, 33.5% for $\alpha = 6$, 35.1% for $\alpha = 7$, 36.6% for $\alpha = 8$, 37.0% for $\alpha = 9$, and 37.6% for $\alpha = 10$), which is consistent with our explanation of experimental logic that intervention on correct answers can in principle have both outcomes. The flip rate is dependent on the relative position of model and human representations compared to the decision boundary, as well as the robustness of the model, and cannot suggest conclusions."

2. Per-subtype paired statistical tests accompanying Fig. S7*, so that the "no consistent failure mode" conclusion is quantitatively supported.*

Response:

Thank you for this valuable suggestion. To address this comment, we perform one-sided paired t-tests over different models ($n = 6$) for NARF > Original under each reasoning subtype. The results show that NARF results are significantly higher than original results for almost all subtypes (Syll_2_affirm: $p = 0.101$; Syll_3_affirm: $p = 0.022$; Syll_2_negate: $p = 0.004$; Syll_3_negate: $p = 0.001$; Tran_2_affirm: $p = 0.004$; Tran_3_affirm: $p = 0.007$; Tran_2_negate: $p = 0.033$; Tran_3_negate: $p = 0.027$). For Syll_2_affirm, p-value is not significant because of the small sample size and small improvement under originally high accuracy for some models, while five of six models show

improvement (up to 12.4%). These results support the conclusion that there is no consistent failure mode to improve the model across reasoning subtypes.

Revision

We have added the results and discussion in Supplementary Information Section S11.1. We quote the revision below:

"To quantify this, we perform one-sided paired t-tests over different models ($n = 6$) for NARF > Original under each reasoning subtype, and the results show that NARF results are significantly higher than original results for almost all subtypes (Syll_2_affirm: $p = 0.101$; Syll_3_affirm: $p = 0.022$; Syll_2_negate: $p = 0.004$; Syll_3_negate: $p = 0.001$; Tran_2_affirm: $p = 0.004$; Tran_3_affirm: $p = 0.007$; Tran_2_negate: $p = 0.033$; Tran_3_negate: $p = 0.027$). For Syll_2_affirm, p-value is not significant because of the small sample size and small improvement under originally high accuracy for some models, while five of six models show improvement (up to 12.4%). These results suggest that there is no subtype that NARF consistently fails to improve."

3. A brief mechanistic discussion of the design choices that plausibly underlie NARF's resistance to catastrophic forgetting.

Response:

Thank you for this valuable suggestion. We agree that a brief discussion would enrich the paper and we have supplemented a discussion in Supplementary Information as quoted below.

Revision

We have added this discussion in Supplementary Information Section S11.5. We quote the revision below:

"This stability is possibly due to both a general property of LLMs that task-specific fine-tuning on limited data typically preserves broad capabilities,

and the design of NARF in which the neural-guided signal enters the objective with directionally-bounded targets or constraints from language labels."

4. An explicit note in the text that Fig. S11 plays the role of a LOSO equivalent subject-robustness analysis

Response:

Thank you for this valuable suggestion. We have added an explicit note in the text to clarify that Fig. S11 serves as a LOSO equivalent subject-robustness analysis, as quoted below.

Revision

We have added this note in Supplementary Information Section S10.4 and S11.6. We quote the revision below:

"This experiment also serves a similar purpose to the Leave-One-Subject-Out (LOSO) analysis, namely examining whether models overfit to specific subjects."

2 To Reviewer 2

We sincerely thank the reviewer for the thoughtful and constructive evaluation of our work, and we are glad that our revision has addressed all concerns.

We have also made some small revisions in response to the comments.

1. It does seem like the alpha parameter is again quite specific to the model though, so this is important to highlight as a sensitivity of the method.

Response:

Thank you for this valuable suggestion. We agree that it is important to highlight this sensitivity and we have supplemented a discussion of sensitivity in Supplementary Information as quoted below.

Revision

We have added a discussion in Supplementary Information Section S10.3. We quote the revision below:

"We also notice that the applicability of representation steering varies for different models, and the α parameter is sensitive and specific to models. Actually, it has been shown that the effectiveness of representation steering depends on the model and is not uniformly reliable [15, 16]. It may be improved by more advanced steering methods in the future work."

2. The second part or your answer gives a key insight into how you think about the modeling, and I think this should be stated explicitly in the introduction: you view the neural data as a tool to improve core ML capabilities, but improved neural alignment is itself not the goal (it is for my research, so I incorrectly assumed it would be for you as well; hence I think good to clarify upfront).

Response:

Thank you for this valuable suggestion. We agree that it is important to clarify this perspective and we have revised the Introduction for clarification, as quoted below.

Revision

We have revised the Introduction to clarify this perspective. We quote the revision below:

"Building on this partial alignment, we introduce an approach to actually enhance Large Language Model (LLM) reasoning using human neural data. In essence, we guide the model's internal representations—that will produce incorrect responses—toward directions informed by the joint structure

of model and brain representations, viewing neural data as a tool for performance improvement rather than treating global model–brain alignment as the ultimate goal."