nature machine intelligence

Supplementary information

https://doi.org/10.1038/s42256-026-01278-w

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

In the format provided by the authors and unedited

Table of Contents

List of Supplementary Figures

List of Supplementary Tables

S1 More details of the fMRI dataset and preprocessing

The functional magnetic resonance imaging (fMRI) dataset (publicly available on OpenNeuro, ID: ds003076) is an accompanying adult dataset to a school-aged children dataset for deductive reasoning [1] and was also part of the data used in [2]. We focus on this adult dataset for its higher behaviour accuracy, enabling robust comparison with and improvement for Large Language Models (LLMs).

The dataset involves two types of deductive reasoning: syllogistic and transitive reasoning. For syllogistic reasoning problems, the premises describe a series of relationships (adjective) among three classes (monosyllabic pseudoword), and the adjective is one of the following: *tall, short, big, small, old, young, fast, slow, brown, red, black, blue, green, white, pink*. For transitive reasoning problems, the premises describe a linear ordering (comparative adjective) of 4 imaginary characters (syllable name), and the comparative adjective is one of the following: *slower, faster, shorter, taller, younger, older, smaller, bigger*.

We use MATLAB, SPM12, and GLMSingle [3] to preprocess fMRI data. We first preprocess the data with slice time correction, motion correction (realigning all functionals to the mean image of the first functional run), normalization to the standard Montreal Neurological Institute (MNI) space by first coregistration to the structural image and then normalization with estimated parameters during segmentation of the structural image (normalized voxel size $2 \times 2 \times 2\,\mathrm{mm}^3$), and smoothing with a 4mm FWHM Gaussian filter.

Then we estimate the response to each problem with GLMSingle, which is an improved method to estimate single-trial fMRI responses based on custom hemodynamic response function (HRF) for each voxel, derivation of a set of noise regressors with cross-validation, and ridge regression for beta estimates [3]. We separate this process for syllogistic and transitive problems because their response time differs in general, and use their average response time as the response time in GLMSingle. For the design matrix, we treat each type of problems with correct or incorrect answers as a condition (used for cross-validation in GLMSingle), resulting in up to $8 \times 2 = 16$ conditions depending on each human subject's behaviour, and extract single-trial responses for each problem with the design matrix.

We extract voxels from the estimated response based on locations of deductive reasoning regions specified in the meta-analysis [4] and a process similar to group-constrained, participant-specific functional localization [5, 6]. Specifically, we first construct a deductive reasoning region mask based on spherical approximations of the coordinates and volume sizes (minimum radius set as 5 mm) of peaks for all deductive arguments, relational arguments (*i.e.*, transitive), and categorical arguments (*i.e.*, syllogistic) specified in [4] (Table 2 and Table 3, we transform the provided Talairach coordinates into the MNI space), yielding a mask involving left inferior frontal gyrus (IFG), left

medial frontal gyrus (MeFG), bilateral middle frontal gyrus (MFG), bilateral precentral gyrus (PG), bilateral posterior parietal cortex (PPC), and left Basal Ganglia (BG). Then we compute individual activation maps within this mask using a contrast condition. Specifically, for each human subject, we first use SPM12 to get the t-values of all voxels considering the contrast: correct_reasoning > premises, where the condition correct_reasoning refers to correctly answered problems (onset time: the start of judgement after presenting the conclusion, duration: response time to the button press), while the condition premises refers to reading premises (onset time: the start of presentation of premises, duration: 6s). The t-values are separately calculated for syllogistic and transitive problems, and we take the max of them for each voxel to select 10% of the most responsive voxels within the mask (with the highest t-values, responsive to either syllogistic or transitive reasoning). These voxels are considered as vectors of brain representations for each human subject. For language and multiple-demand (MD) regions, their masks are available from https://www.evlab.mit.edu/resources, which are widely used in previous works [6], and we follow the same procedure as deductive reasoning regions.

S2 More details of the test datasets

S2.1 Generated reasoning problems

Syllogistic and transitive problems. The generated questions follow the structure of questions in the fMRI dataset, while introducing new lexical items and premise numbers. The monosyllabic pseudowords and syllable names are randomly generated and different from the fMRI dataset, while the adjectives are randomly chosen from a list – for syllogistic reasoning, the list of adjectives is: *strong, weak, bright, dark, warm, cold, soft, hard, smooth, rough, quiet, loud, sharp, dull, heavy, light, wide, narrow, deep, shallow, clean, dirty, fresh, stale, rich, poor, brave, cowardly, cheerful, sad, proud, humble*; for transitive reasoning, the list of comparative adjectives is: *brighter, darker, richer, poorer, stronger, weaker, quieter, louder, cleaner, dirtier, fresher, staler, heavier, lighter, smoother, rougher, warmer, colder, wider, narrower, deeper, shallower, sharper, duller, softer, harder, prouder, humbler, braver, calmer, sadder, happier*, all distinct from the original fMRI stimuli.

For three-premise problems, we maintain the original fMRI dataset's eight conditions (true/false × affirmative/negation × 2/3 required premises). For problems with more premises (4-6), we consider all premises to be used and organize problems into four conditions (true/false × affirmative/negation). We generate the same number of questions for all groups, enabling a more balanced measurement than the fMRI dataset.

During validation and testing, we enumerate all possible premise permutations ($N!$ variants for N premises). For example, for three premises (1) *A is stronger than B*, (2) *B is stronger than C*, (3) *C is stronger than D*, there will be six possible orders, such as (1) *C is stronger than A*, (2) *B is stronger than C*, (3) *A is stronger than B*. This controls the overfitting to the fixed pattern

of linear orders and comprehensively evaluates the deductive reasoning ability of models that may differ for different premise orders [7].

For the validation set, we consider problems with three premises and generate 10 problems for each condition, resulting in $10 \times 8 \times 2 = 160$ problems and finally $160 \times 3! = 960$ validation data. For the test set with three/four/five/six premises, we generate 100/50/20/10 problems for each group, leading to 1600/400/160/80 problems and finally 9600/9600/19200/57600 test data.

Propositional problems. We generate propositional problems with three types: modus ponens, modus tollens, and disjunction elimination [4], each with true or false conclusions. For modus ponens with the correct conclusion, the two premises are in the format: (1) *If pseu1 is adj1, then pseu2 is adj2*, (2) *pseu1 is adj1*, and the conclusion is *pseu2 is adj2*, where *pseu1* and *pseu2* are two randomly generated pseudowords and *adj1* and *adj2* are two adjectives randomly chosen from: *strong, weak, bright, dark, warm, cold, soft, hard, smooth, rough, quiet, loud, sharp, dull, heavy, light, wide, narrow, deep, shallow, clean, dirty, fresh, stale, rich, poor, brave, cowardly, cheerful, sad, proud, humble*. For modus ponens with the incorrect conclusion, the conclusion is *pseu2 is not adj2*. For modus tollens, the second premise is replaced by *pseu2 is not adj2* and the conclusion is *pseu1 is not adj1* or *pseu1 is adj1* correspondingly. For disjunction elimination, the two premises are in the format: (1) *pseu1 is adj1 or pseu2 is adj2*, (2) *pseu1 is not adj1*, and the conclusion is *pseu2 is adj2* or *pseu2 is not adj2*. We generate 100 problems for each condition, resulting in 600 problems.

S2.2 FOLIO dataset

The FOLIO dataset [8] is a first-order logical inference dataset for reasoning in natural language. Each question consists of multiple premises and a hypothesis. An example of the dataset is:

Premises:

(1) The Blake McFall Company Building is a commercial warehouse listed on the National Register of Historic Places.

(2) The Blake McFall Company Building was added to the National Register of Historic Places in 1990.

(3) The Emmet Building is a five-story building in Portland, Oregon.

(4) The Emmet Building was built in 1915.

(5) The Emmet Building is another name for the Blake McFall Company Building.

(6) John works at the Emmet Building.

Hypothesis:

A five-story building is built in 1915.

We follow the FOLIO-wiki-curated version [9], a refined subset removing problematic instances. The answers to the hypothesis can be "True", "False", or "Unknown", and we only consider questions with "True" or "False" answers (*i.e.*, FOLIO-wiki-curated-TF) in order to align with the setting of our training data, with a total number of 315 questions.

S3 More details of model prompting

The prompt for syllogistic reasoning is:

"*You are an expert in performing logical reasoning. I will present three premises and one conclusion each round. The premises describe a series of relationships among monosyllabic pseudowords and adjectives. Assume all premises are True. You should use deductive reasoning to determine if the conclusion can be drawn from the premises logically. Answer 'True' if the conclusion can be drawn logically; otherwise you should answer 'False'. Only answer 'True' or 'False' at the beginning of the response.*"

For transitive reasoning, the description is slightly different:

"*The premises describe relationships among imaginary characters with comparative adjectives.*"

For problems with more premises, the description correspondingly changes to the number of premises. And for propositional reasoning, the corresponding description is:

"*I will present two premises and one conclusion each round. The premises describe logical arguments.*"

After the model's round saying "*Understood. I will do my best to determine if the conclusion can be logically drawn from the given premises, and only answer 'True' or 'False' at the beginning of the response.*", the premises and conclusion are presented, *e.g.*, "*Premises:\n 1. All blons are pink.\n 2. All pink things are young.\n 3. Ken is a blon.\n Conclusion:\n Ken is pink.*" Then the model is required to output the answer at its round in the chat template.

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.*"

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

S4 More details of ceiling calculation for neural predictivity

For neural predictivity calculation, the raw scores are divided by an estimated upper-bound ceiling based on predictivity between human neural activations. This is due to the intrinsic noise in biological measurements, so a ceiling value reflecting the explainable variance in neural activations is usually adopted for brain scores. We calculate the ceiling value in a similar way as the established

practice [6]. First, we subsample the n-participant data into all possible combinations of s participants ($s \in [2, n]$). For each subsample, we designate a target participant in the subsample and predict its neural signals from the remaining participants using ridge regression that matches the main analysis pipeline. Then we calculate the highest possible estimate for each voxel of each human subject by extrapolating the score to infinitely many participants through fitting the equation $v = v_0 \times \left(1 - e^{-\frac{s}{\tau_0}}\right)$ for each voxel, where v represents the score and v_0 and τ_0 are parameters to fit. Finally, the ceiling score takes the average of v_0 for all voxels of all human subjects. The calculated ceiling values are 0.257, 0.299, 0.28 for deductive reasoning regions considering all problems, syllogistic problems, and transitive problems; 0.248, 0.287, 0.26 for MD regions; and 0.253, 0.293, 0.27 for language regions. They are comparable to previous ceiling values of language regions in the language comprehension task (0.32 and 0.20 for two fMRI datasets) [6].

S5 More explanations of NARI and NARF

Both neural activation guided representation intervention (NARI) and fine-tuning (NARF) leverage the joint structures of model representations and human neural activations. Specifically, consider the ridge regression from the space of model representations to the space of neural signals, let $\mathbf{X}$ and $\mathbf{Y}$ denote the data of model representations and neural activations (columns as samples), the weight $\mathbf{W}$ and intercept $\mathbf{b}$ of ridge regression ideally approximate the closed-form solution:

$$\begin{aligned} \mathbf{W} &= \mathbf{Y}_c^\top \mathbf{X}_c \left(\mathbf{X}_c^\top \mathbf{X}_c + \lambda \mathbf{I}\right)^{-1}, \\ \mathbf{b} &= \overline{\mathbf{y}} - \mathbf{W}\overline{\mathbf{x}}, \end{aligned} \tag{S1}$$

where $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are sample means, $\mathbf{X}_c = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^\top$ and $\mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\overline{\mathbf{y}}^\top$ are centered data, and λ is the ridge coefficient.

For the objective $S = \mathrm{Sim}(\mathbf{W}\mathbf{x}, \mathbf{y})$ with the data point $\mathbf{x}$ and $\mathbf{y}$, when Sim is taken as cosine similarity or the negative of mean-square-error (MSE), the

gradient of $\mathbf{x}$ is:

$$\begin{aligned}
\nabla_{\mathbf{x}} S_{cos} &= \frac{1}{\|\mathbf{W}\mathbf{x}\|\|y\|}\mathbf{W}^\top\mathbf{y} - \frac{\langle \mathbf{W}\mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{W}\mathbf{x}\|^3\|\mathbf{y}\|}\mathbf{W}^\top\mathbf{W}\mathbf{x}, \\
&= \frac{1}{\|\mathbf{W}\mathbf{x}\|\|y\|}\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^\top\mathbf{Y}_c\mathbf{y} \\
&\quad - \frac{\langle \mathbf{W}\mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{W}\mathbf{x}\|^3\|\mathbf{y}\|}\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^\top\mathbf{Y}_c\mathbf{Y}_c^\top\mathbf{X}_c\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{x}, \\
\nabla_{\mathbf{x}} S_{-mse} &= -\mathbf{W}^\top\mathbf{W}\mathbf{x} + \mathbf{W}^\top\mathbf{y} \\
&= -\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^\top\mathbf{Y}_c\mathbf{Y}_c^\top\mathbf{X}_c\left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{x} \\
&\quad + \left(\mathbf{X}_c^\top\mathbf{X}_c + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_c^\top\mathbf{Y}_c\mathbf{y}.
\end{aligned} \tag{S2}$$

Eq. (S2) shows that the direction (gradient) will be induced by $\mathbf{X}_c^\top\mathbf{X}_c$, $\mathbf{Y}_c\mathbf{Y}_c^\top$, and $\mathbf{X}_c^\top\mathbf{Y}_c$, which are the structure of model representations, the structure of human neural activations, and the interactive structure of model and neural representations over all considered reasoning problems. This is a joint effect of model and neural representations, which varies among different models and human subjects (Sec. S9). For the baseline that replaces the fMRI data with random signals, it eliminates the effect of human neural activations ($\mathbf{Y}_c\mathbf{Y}_c^\top$ and $\mathbf{Y}_c\mathbf{y}$ are expected to be a constant matrix/vector), isolating the effect of model representational structure alone. Therefore, this ablation confirms the distinct contribution of neural guidance to representation improvement.

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.

Note that we calculate the gradients using the objective $S = \mathrm{Sim}(\mathbf{W}\mathbf{x}, \mathbf{y})$ rather than the intercept-inclusive form $S_{w/\ icpt} = \mathrm{Sim}(\mathbf{W}\mathbf{x} + \mathbf{b}, \mathbf{y})$, 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 induces 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 $\nabla_{\mathbf{x}} S_{w/\ icpt} = -\mathbf{W}^\top\mathbf{W}\mathbf{x} - \mathbf{W}^\top\mathbf{b} + \mathbf{W}^\top\mathbf{y}$, while the gradient for the intercept-exclusive form is $\nabla_{\mathbf{x}} S = -\mathbf{W}^\top\mathbf{W}\mathbf{x} + \mathbf{W}^\top\mathbf{y}$, meaning $\nabla_{\mathbf{x}} S - \nabla_{\mathbf{x}} S_{w/\ icpt} = \mathbf{W}^\top\mathbf{b}$ so that the gradient ascent direction additionally incorporates 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].

S6 More implementation details of NARI and NARF

In experiments, Sim is taken as cosine similarity by default, which empirically shows slightly better performance (Fig. 4f). Overall, cosine similarity is similar to the negative of MSE because they produce similar gradient components (Eq. (S2)). For the combination of NARF and language supervision, we take Sim as the negative of MSE for Qwen2-7B-Instruct and Llama2-7B-Chat to achieve slightly higher performance. In practice, to align with the common gradient descent framework, we minimize $1 - \mathrm{CosSim}(x, y)$ or $\mathrm{MSE}(x, y)$ and apply gradient descent.

For Neural Activation guided Representation Intervention (NARI), we apply projected gradient descent to modify representations based on the objective. Specifically, given the gradient $\nabla_{\mathbf{r}^l} S$, we first perform gradient descent with the Adam optimizer to obtain the updated representation $\mathbf{r}^{l'}$ and the difference $\Delta\mathbf{r}^l = \mathbf{r}^{l'} - \mathbf{r}^l$, and then project the difference in the specified range $\alpha\sigma^l$, *i.e.*, $\Delta\mathbf{r}^l_p = \frac{\min\left(\|\Delta\mathbf{r}^l\|, \alpha\sigma^l\right)}{\|\Delta\mathbf{r}^l\|}\Delta\mathbf{r}^l$, to get the projected representation $\mathbf{r}^{l'}_p = \mathbf{r}^l + \Delta\mathbf{r}^l_p$. We perform multiple gradient descent steps and verify if the modification direction enables the model to correct its answer for every 5 steps. The verification performs intervention for representations of all considered layers, *i.e.*, the representation of each layer will add the difference $\Delta\mathbf{r}^l_p$ during forward propagation, and the modification can be accumulated across layers. Once the intervention corrects the answer, we stop the steps and treat it as successful. The maximum gradient descent steps are set as 200 for non-thinking models and 50 for the thinking model DeepSeek-Distill-Qwen-1.5B. The successfully intervened representation will be saved for the vanilla NARF method. For the baseline of random signals, it replaces fMRI data with randomly generated data and follows the same process. For the baseline of random directions, we sample multiple directions equivalent to the number of directions of NARI, perform grid search for the scale of the direction from σ^l to $\alpha\sigma^l$, and treat the intervention as successful as long as the answer can be corrected during grid search.

For the general direction of NARI, we take the average of successful intervention directions $\Delta\mathbf{r}^l_p$ for all questions, with $\alpha = 10$ for non-thinking models and $\alpha = 5$ for DeepSeek-Distill-Qwen-1.5B. For DeepSeek-Distill-Qwen-1.5B, we only consider the nearest successful intervention point for each question.

The scale γ is searched between 0.1 and 1.0 with the interval 0.1 for Qwen2-1.5B-Instruct, Llama2-7B-Chat, and Phi-4-mini-Instruct, between 0.02 and 0.2 with the interval 0.02 for Llama3-8B-Instruct, and between 0.5 and 5.0 with the interval 0.5 for DeepSeek-Distill-Qwen-1.5B.

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 like other steering methods [11]. For the per-instance direction calculation of NARI, the gradient ascent for representations only has 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 come 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.

We primarily followed the prior practice of inference-time intervention [11, 12] to perform 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.

For the combination of NARF and language labels, we start training from the original model by default. For Llama2-7B-Chat, we find that starting from the NARF-trained model leads to higher performance and report this result. We do not add the regularization term in this setting. For Qwen3-4B, Gemma2-9B-it, and Llama3.3-70B-Instruct, we additionally filter three human subjects whose overall success rates for the intervention experiments are below 60% (see Fig. S5c), and use the remaining human neural signals for experiments. 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.

S7 More neural predictivity results and discussions

Layer-wise neural predictivity. Fig. S1 presents the predictivity scores of different layers and modules (attention module output or layer output) for various models. Results show that the middle layers are generally more predictive, which may accord with the common practice to analyze middle layers because they are likely to contain interesting and abstract features [13]. Results also show that the attention module outputs usually have slightly higher scores than layer outputs, aligning with [14] that emphasize the importance of analyzing "transformations" (*i.e.*, the output of attention heads).

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 exhibit 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.

S8 Comparison of neural predictivity results between instruction-tuned and base models

We compare the neural predictivity results between Qwen2-7B-Instruct and Qwen2-7B (base model) as well as Llama2-7B-Instruct and Llama2-7B (base model) in Fig. S2. For the base model without chat format, we present the premises and conclusion directly after the task prompt and add "Answer:\n" as the generation prompt. Representations are also extracted at the last token. Results show that the base models are similar to the instruction-tuned models, considering the comparable predictivity scores, the same specificity over brain regions, and the similar discrepancy of predictivity under various reasoning type settings. The layer-wise score distributions also show similar trends.

S9 Structures of models' and human brains' representations

As analyzed in Sec. S5, the proposed NARI and NARF are induced jointly by the structures of model and human neural representations. We visualize such structures (the Gram matrices $\mathbf{X}_c^\top \mathbf{X}_c$ and $\mathbf{Y}_c \mathbf{Y}_c^\top$ of centered data as described in Sec. S5) for different models and human subjects in Fig. S3, considering 34 transitive reasoning problems. Results show that the structures vary among different models, layers of the same models, and different human subjects. Therefore, the joint effect of them in NARI/NARF would be sample-specific, motivating our multi-subject integration approach to capture robust improvement signals.

S10 Extended NARI results

S10.1 Model-specific representation intervention results

The reported results of NARI in the main text (Fig. 4) are the summary of six models. We further report the detailed results of individual models in Fig. S4, showing that the conclusion is consistent among all models.

S10.2 More analysis results of NARI

In this section, we provide more analysis results to show the importance of incorporating the systematic shift term in NARI as discussed in Sec. S5 and the variance among human subjects.

First, the systematic shift term proves essential for effective representation steering. Fig. S5a-b demonstrate its superiority across all intervention ranges, with three key observations: 1) Random signal baselines also show improved success rates when incorporating the shift term (which is a shift between transformed model representations and zero), indicating its role in correcting model representation biases; 2) The shift term alone (Rand. Sig. w/ fMRI Mean condition in Fig. S5a) yields negligible effects, confirming that neural signal structure remains indispensable; 3) optimal performance requires the joint effect between neural structure and systematic shift terms.

Second, human subject variability impacts intervention outcomes. Fig. S5c reveals substantial variation in subject-specific success rates of six models (range: 52-86%), where the subject-specific success rate is computed as the percentage of successfully intervened reasoning problems among all problems where the model answers incorrectly while the particular human subject is correct (the problems are different for distinct human subjects). The variability aligns with the representational diversity shown in Sec. S9, indicating that the joint effect of structures for intervention is sample-specific.

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.

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, 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.

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. This experiment also serves a similar purpose to the Leave-One-Subject-Out (LOSO) analysis, namely examining whether models overfit to specific subjects. 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. These results verify the robustness of our method.

S11 Extended NARF results

S11.1 More results on performance across reasoning subtypes

In Fig. S7, we provide more results of each model's performance across reasoning subtypes after fine-tuning. Subtypes are defined by the reasoning type and condition (Methods, Sec. S2.1): for each reasoning type (syllogistic/transitive), conclusion form (affirmative/negation), and premise requirement (2/3 to make judgement), we gather the results of problems with "true" or "false" conclusions, leading to 8 subtypes such as Syll_2_affirm. Fig. S7 shows that the improvement of NARF compared to original models or language-label-only fine-tuned models is consistent for most subtypes on the test data. 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.

S11.2 Ablation analysis of NARF when combined with language supervision

As analyzed in Sec. S5, NARF is also induced by the joint effect of model and neural representational structures. To analyze the effect of the components, we perform ablation studies by replacing the human fMRI signals with random data (random signal induced representation fine-tuning (RSRF)) similar to the baseline in NARI experiments, or with one-hot encodings of label signals (label signal induced representation fine-tuning (LSRF)). RSRF eliminates the effect of human neural activations and is only induced by the model representational structure (Sec. S5). LSRF introduces label signals and is induced by the joint effect of model representational structure and labels. As shown in Fig. S8, while RSRF+Label also improves label-only fine-tuning due to the information in model representations, NARF yields significantly larger gains, indicating that human neural structures provide unique improvement signals beyond model-internal patterns, and the improvement is also a joint effect of model and neural representations. Meanwhile, LSRF+Label does not show improvement over RSRF+Label (sometimes even has negative impacts on label fine-tuning), indicating that the structure of labels cannot effectively guide representations for generalization to new problems, highlighting the effectiveness of the structures of human neural activations.

S11.3 Similarity metrics with human neural activations

As we optimize models' representations toward directions encouraging a higher similarity metric with human neural activations, we further analyze the similarity metrics after fine-tuning. We first analyze the similarity metric calculated as the objective in NARF (*i.e.*, $S = \mathrm{Sim}(\mathbf{Wx}, \mathbf{y})$ that removes the intercept term) and we take the average of similarity among all considered layers. As shown in Fig. S9a, NARF and NARF+Label largely increase this similarity metric, while language-label-only training hardly realizes it, indicating that they are different directions. NARF+Label has much higher similarity because it directly optimizes this objective, while NARF only encourages representations to the nearest successfully intervened point induced by human neural activations. We also analyze the similarity metric following the common predictivity calculation (based on $\mathbf{Wx} + \mathbf{b}$ and the Pearson r correlation metric among data samples). As shown in Fig. S9b, NARF does not show consistent increase/decrease, indicating that the directions are different from optimizing this predictivity score. 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.

S11.4 More visualizations of reasoning trajectories

We present more visualizations of the layer-wise reasoning trajectories for various models in Fig. S10, demonstrating that compared with language-label-only fine-tuning, NARF+Label consistently improves the separation between paths for correct and incorrect logics, leading to higher performance.

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 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. 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.

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. This experiment also serves a similar purpose to the LOSO analysis, namely examining whether models overfit to specific subjects. As shown in Fig. S11c, it is also similar for NARF+Label that including high-performing subjects leads to 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. These results verify the robustness of our method.

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".

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.

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.

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 × 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 for which we find that 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%), this gap is likely due to the increased difficulty introduced by the modality shift, reflecting a boundary condition under which our method remains effective. In addition, Fig. S14d further shows that the general direction can 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.

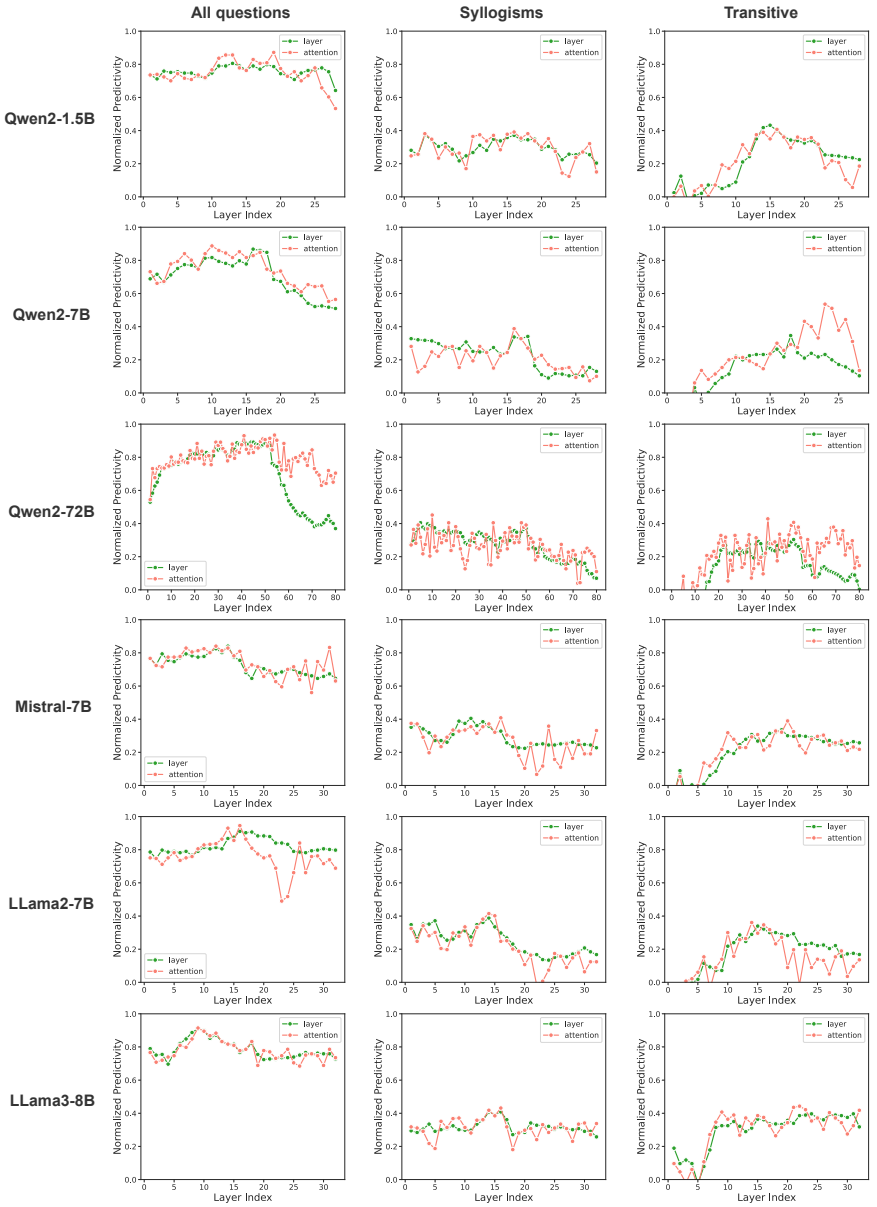

Fig. S1 Layer-wise Neural Predictivity Results of Different Models. For all models, the middle layers generally have higher scores, and attention module outputs generally have slightly higher scores than layer outputs. Models are all instruction-tuned by default, with their names shortened in the figures.

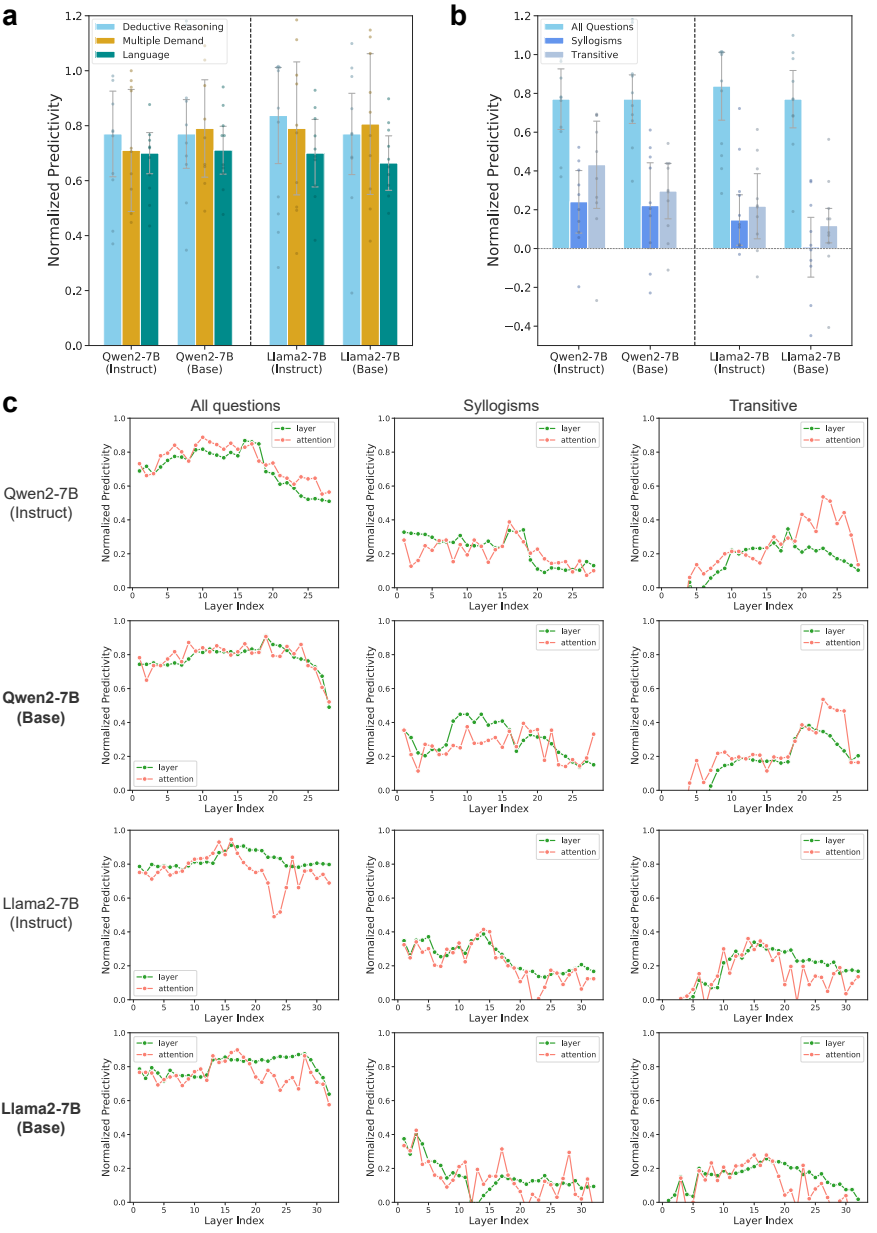

Fig. S2 Comparison Results of Neural Predictivity between the Instruction-tuned and Base Models. (a) Neural predictivity results for different brain regions considering all reasoning problems. Data are presented as median ± median absolute deviation (m.a.d.) considering inter-subject variability for each model's results ($n = 10$), with individual data points overlaid (the same for (b)). (b) Neural predictivity results for different reasoning types. (c) Layer-wise neural predictivity results.

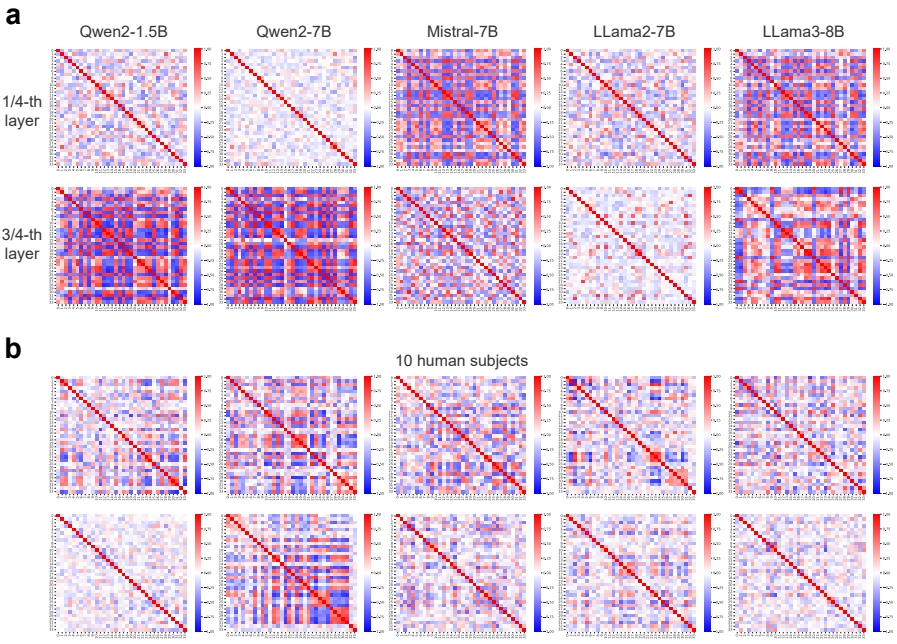

Fig. S3 Visualization of the Structures of Model Representations and Human Neural Activations. The structures are the Gram matrices of centered representations for 34 transitive reasoning problems. (a) The structures of different models' representations at the 14-th layer and 34-th layer, showing model-dependent and layer-dependent variations. (b) The structures of neural representations from 10 human subjects, demonstrating individual-specific variations.

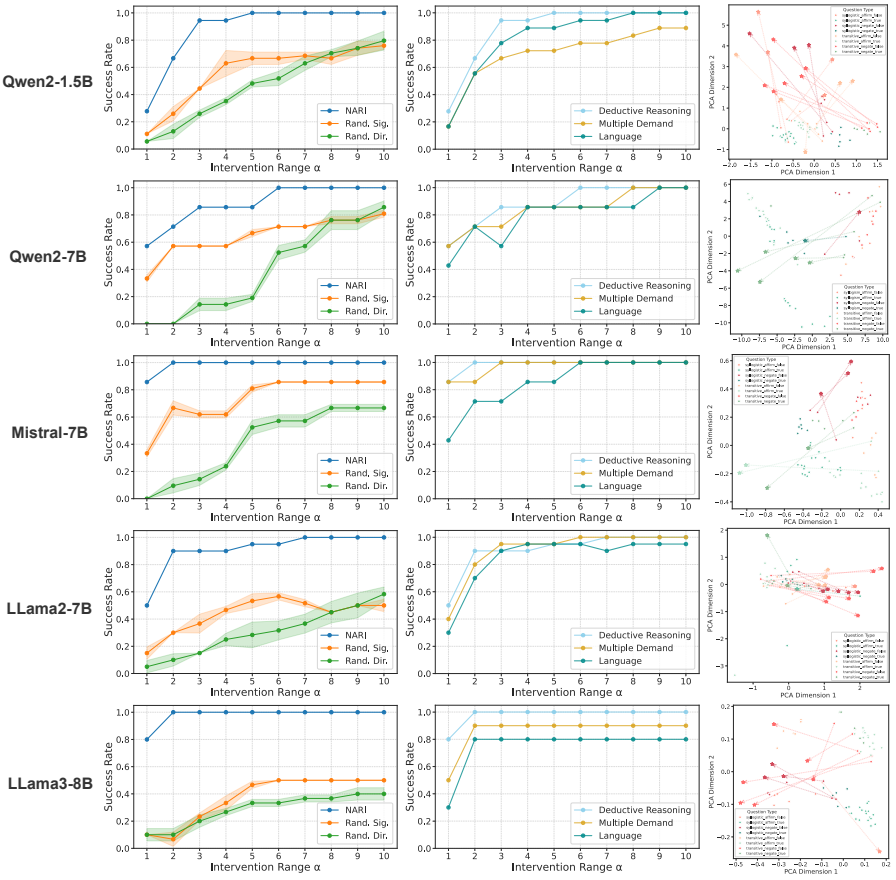

Fig. S4 Model-specific Representation Intervention Results. The first column contains comparison results with the baselines of random signals and random directions. For baselines, data are presented as mean ± std over three runs. The second column contains comparison results of different brain regions. The third column presents visualizations of the intervention processes.

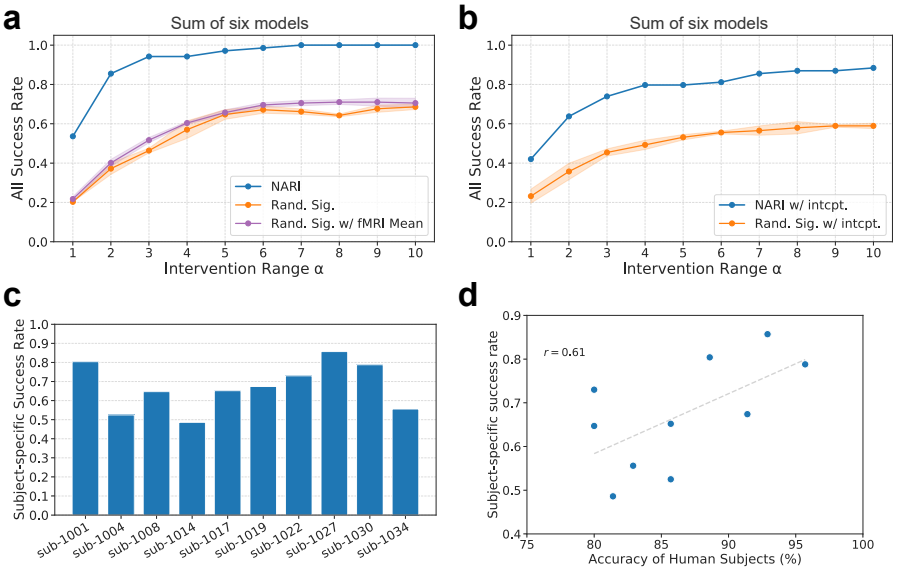

Fig. S5 More Analysis Results of Representation Intervention Experiments. (a) Results with the additional systematic shift term during direction calculation. "Rand. Sig. w/ fMRI Mean" refers to replacing fMRI signals with random signals whose mean vector keeps the same as fMRI signals. For baselines, data are presented as mean ± std over three runs (the same for (b)). (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.

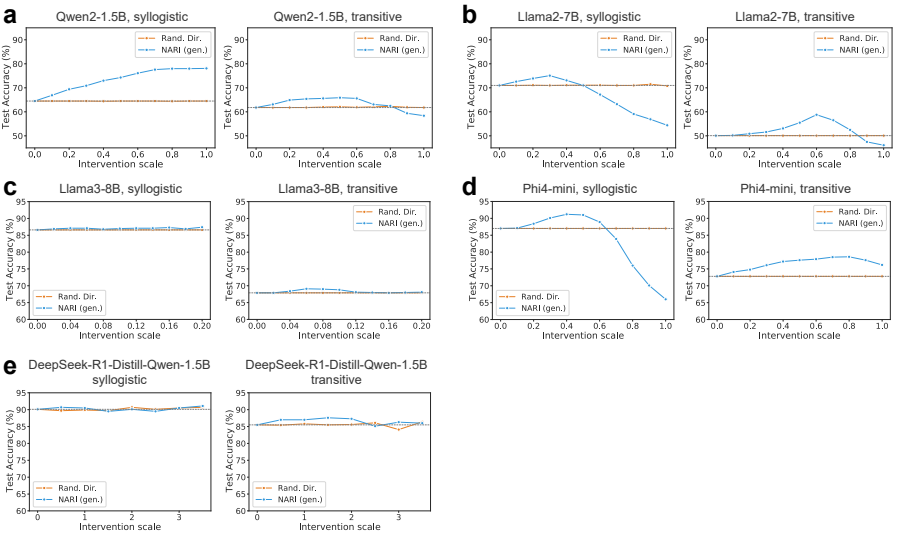

Fig. S6 Analysis Results of General Direction Intervention Experiments. Test accuracy of NARI (gen.) and random direction with different scales for various models.

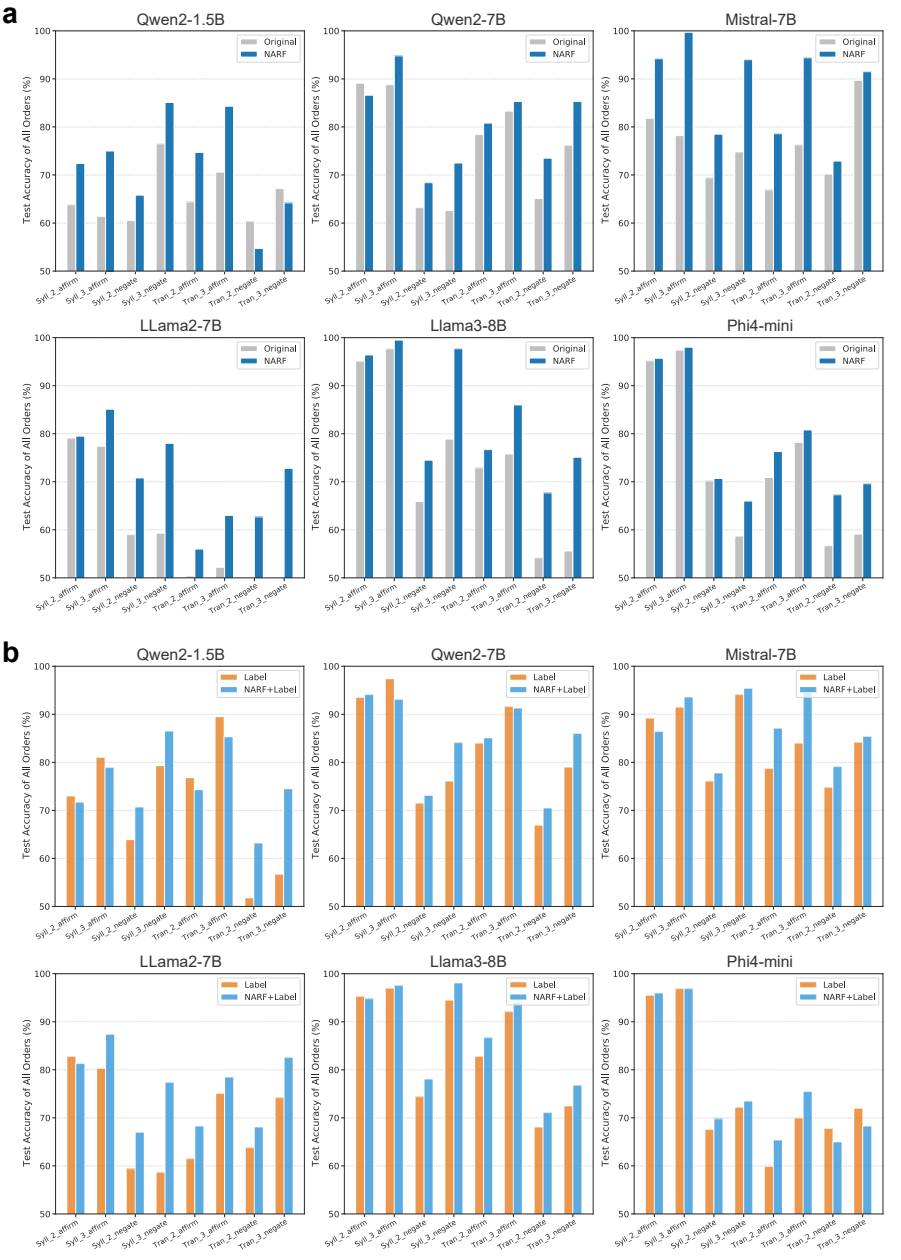

Fig. S7 More Results on Performance across Reasoning Subtypes after Fine-tuning. (a) Comparison results between the original and NARF fine-tuned models. (b) Comparison results between the language-label-only fine-tuned and NARF+Label fine-tuned models.

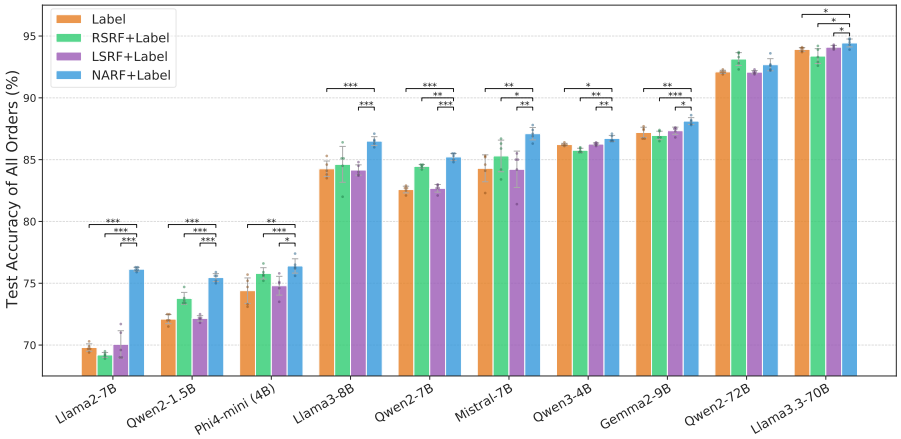

Fig. S8 Ablation Analysis of NARF when Combined with Language Supervision. Statistical tests are reported based on two-sided paired t-test over five runs of experiments under same random seeds. RSRF replaces fMRI signals with random data, which eliminates the effect of human neural activations and is 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.

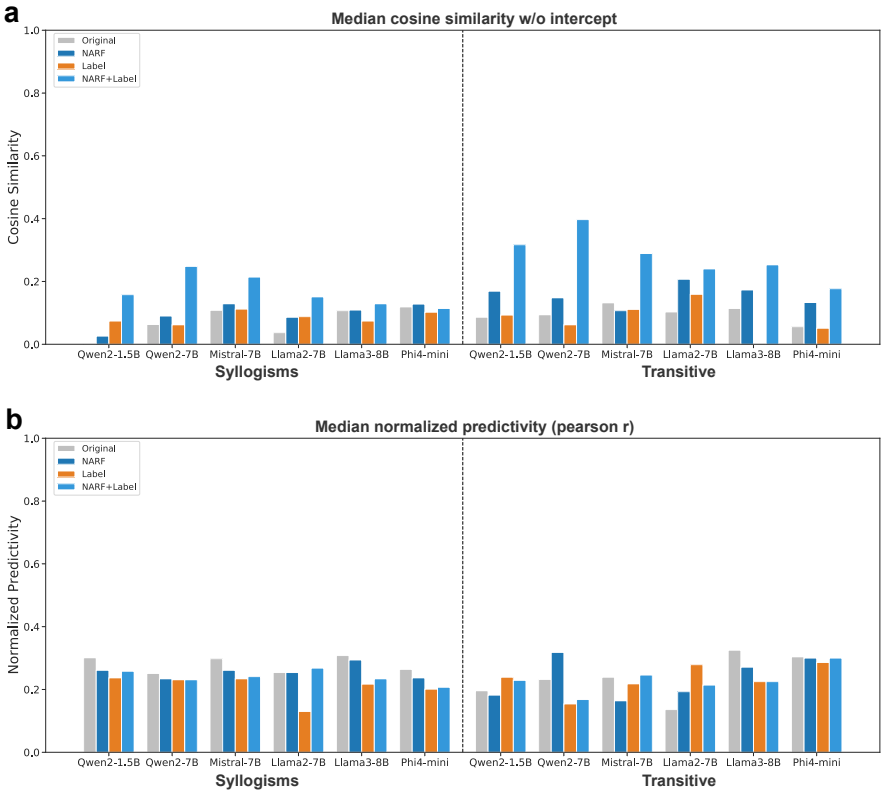

Fig. S9 Results of Similarity Metrics with Human Neural Activations after Fine-tuning. (a) Results of the intercept-excluded similarity metric (our objective), which removes the intercept term and is based on cosine similarity. (b) Results of the similarity metric following the predictivity calculation,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

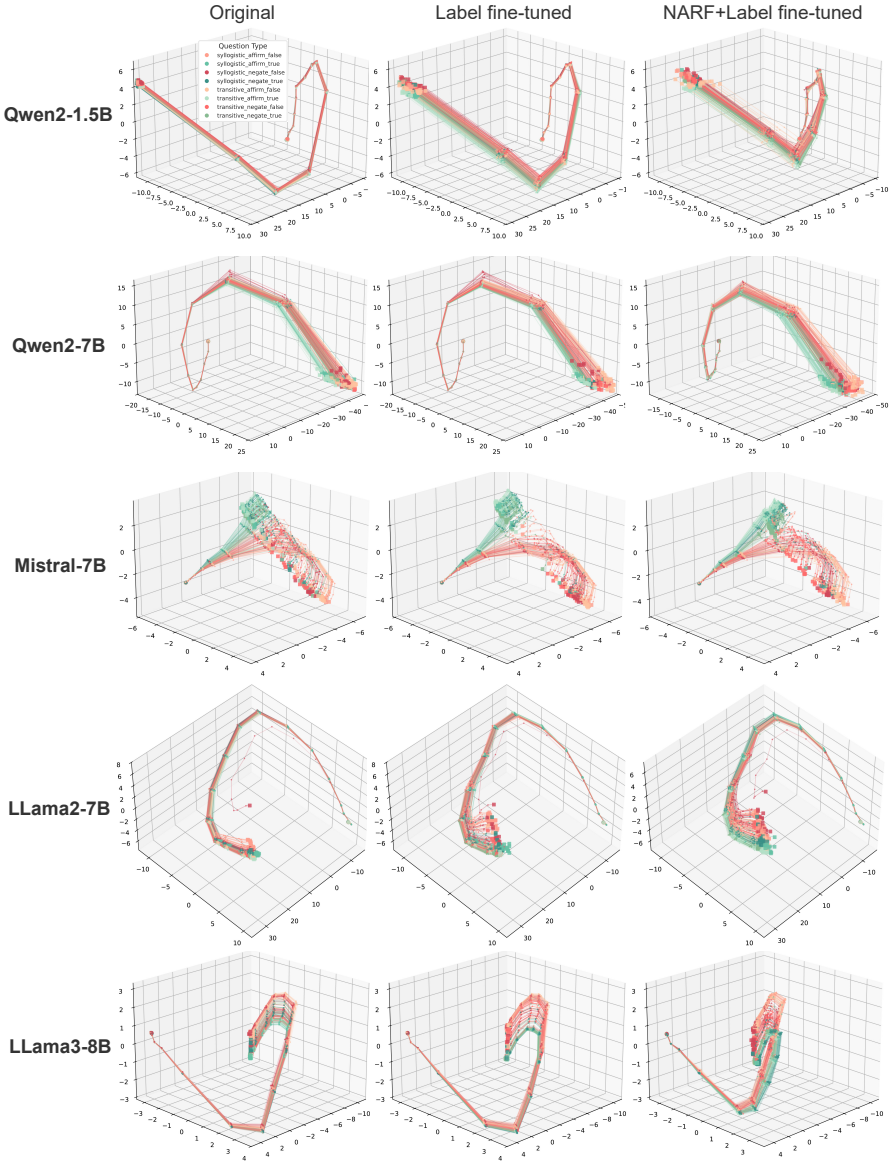

Fig. S10 More Visualization Results of the Layer-wise Reasoning Trajectories in the Representation Space. The trajectories are layer outputs from the 1/4-th layer to the 3/4-th layer, projected into the three-dimensional space based on principle component analysis (PCA).

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

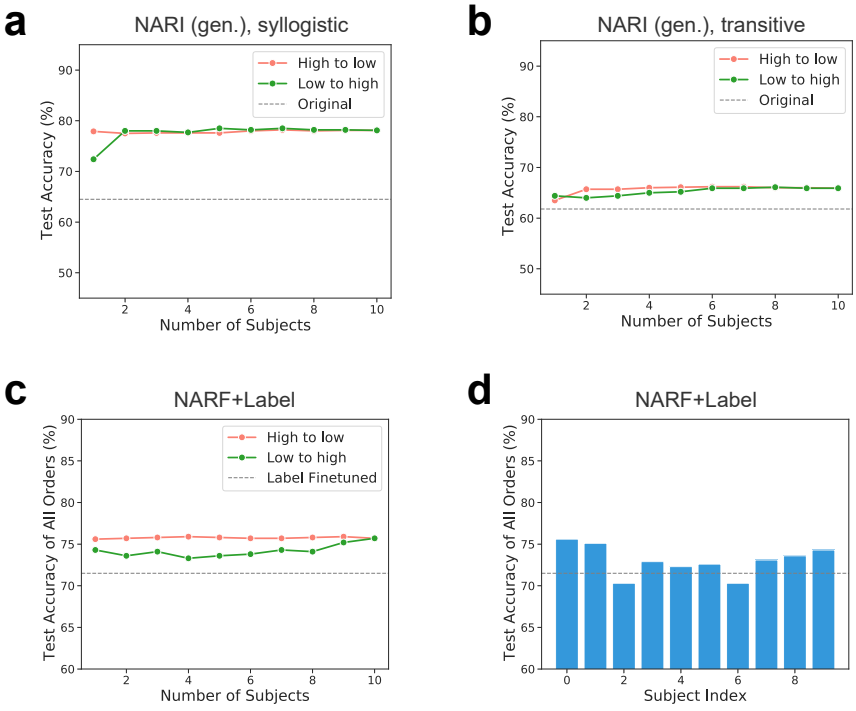

Fig. S11 Analysis Results of the Influence of Different Human Subjects on NARI (gen.) and NARF+Label. (a-b) Test performance of NARI (gen.) on syllogistic and transitive problems with varying number of subjects in two directions: from high-performing to low-performing subjects and vice versa. (c) Test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

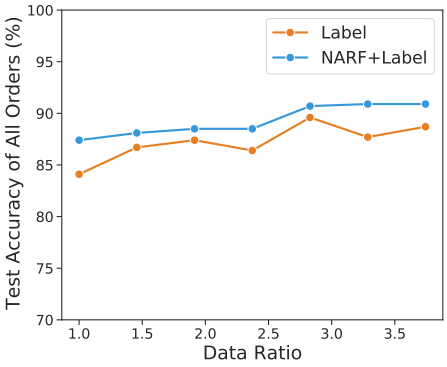

Fig. S12 Results of Combining NARF with Synthetic Data. Data ratio refers to the ratio of training data (problems from the fMRI dataset plus new synthetic data) over original problems from the fMRI dataset.

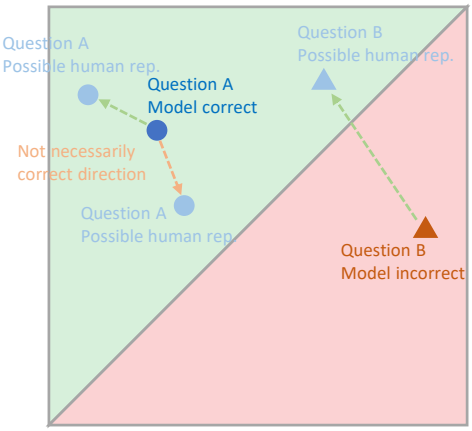

Fig. S13 Illustration of experimental logic. We only intervene on problems that models are incorrect because for those that models are already correct, human representations do not necessarily provide correct direction.

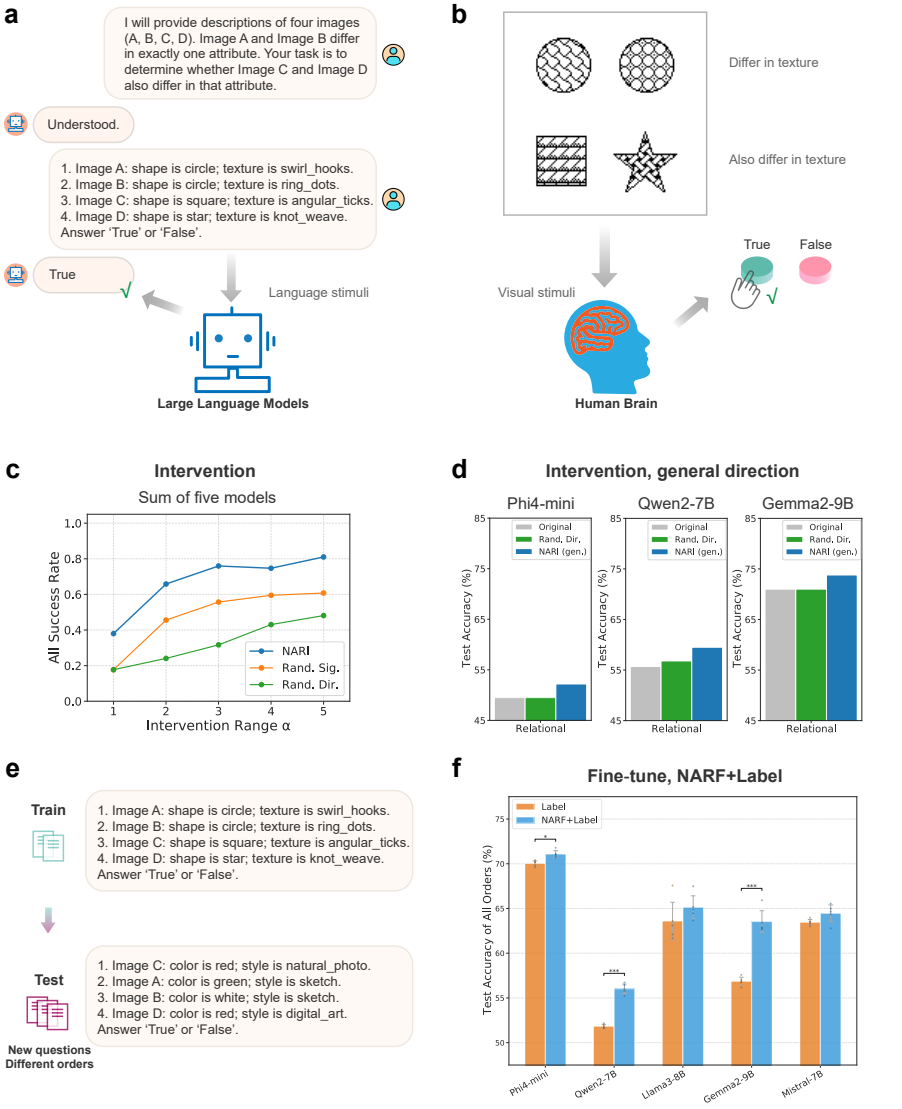

Fig. S14 Results on the Relational Processing Task from the HCP Dataset. (a-b) Illustration of the task for LLMs and humans. The task probes relational processing by comparing relations from two pairs of images, judging whether bottom images differ in the attribute that top image differ in. LLMs process 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 (mean±std) of various models for language-label-only and NARF+Label fine-tuning. Statistical tests are reported based on two-sided paired t-test over five runs of experiments under same random seeds.

References

[1] MN Lytle, J Prado, and JR Booth. A neuroimaging dataset of deductive reasoning in school-aged children. *Data in Brief*, 33:106405, 2020.

[2] Jerome Prado, Rachna Mutreja, and James R Booth. Fractionating the neural substrates of transitive reasoning: Task-dependent contributions of spatial and verbal representations. *Cerebral Cortex*, 23(3):499–507, 2013.

[3] Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, 11:e77599, 2022.

[4] Jérôme Prado, Angad Chadha, and James R Booth. The brain network for deductive reasoning: a quantitative meta-analysis of 28 neuroimaging studies. *Journal of Cognitive Neuroscience*, 23(11):3483–3497, 2011.

[5] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010.

[6] 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.

[7] Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *International Conference on Machine Learning (ICML)*, 2024.

[8] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. Folio: Natural language reasoning with first-order logic. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[9] Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*, 2023.

[10] Nathan Cloos, Moufan Li, Markus Siegel, Scott L Brincat, Earl K Miller, Guangyu Robert Yang, and Christopher J Cueva. Differentiable optimization of similarity scores between models and brains. In *International Conference on Learning Representations (ICLR)*, 2025.

[11] 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)*, 2023.

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

[13] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel

Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/, 2024. Transformer Circuits Blog.

[14] Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1):5523, 2024.

[15] 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.

[16] 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.

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

[18] 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.