BEYOND TASK-CONDITIONED IMPORTANCE: HEAD-RESOLVED OUTPUT JACOBIANS SEPARATE BROAD RESPONSE FROM BEHAVIORAL RELEVANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability in explainable AI is often quantified by the importance of an internal component by how much it 'contributes' to a given prediction/task/behavior. Here we ask the orthogonal question: can an unimportant component for one behavior be highly gainful in other directions of output space? We formulate a head-resolved framework that disentangles gross downstream responsivity, task-guided sensitivity/alignment, and behavioral visibility at the prediction head. This is one of the first systematic comparison of head-output-to-vocabulary Jacobian response with both task-directed sensitivity and behavioral readouts across multiple language models. We study GPT-2 Small, Pythia-410M, and GPT-Neo-125M using output Jacobians, activation patching, and direct logit attribution. The clearest case occurs in Pythia-410M: its two most broadly responsive heads rank only 263rd and 374th in behavioral visibility on indirect object identification. Finite interventions closely match the Jacobian-predicted slopes (approximately 1% relative error at the tested scales), while normalized output gain remains comparable to high-response controls matched with sensitivity. We find that their finite-head gain also varies when we change the task contrasts. This demonstrates that poor importance of task-conditioned task need not correlate with weak local responses downstream. Gross gain, task-guided alignment, and behavioral visibility are therefore orthogonal views of a given model component, rather than synonyms for importance.

1 INTRODUCTION

Modern interpretability methods often ask which model components matter for a particular prediction, task, or behavior. Causal tracing, activation patching, ablation, direct logit attribution, and gradient-based methods provide useful answers to this task-conditioned question (Vig et al., 2020; Meng et al., 2022; Olsson et al., 2022; Wang et al., 2023; Conmy et al., 2023; Syed et al., 2024). A low score, however, does not by itself establish that a component has little ability to affect other output directions.

We study this distinction for individual transformer attention heads at the final prediction position. For each head, we measure the local Jacobian from its output to the full vocabulary-logit vector and compare that broad response with task-directed sensitivity, activation patching, and direct logit attribution. At first order, $\Delta\ell \approx \mathbf{J}_h \Delta\mathbf{z}_h$: the Jacobian describes local response geometry, while realized behavior also depends on the activation change actually produced by the computation. A head can therefore have high local gain yet weak task alignment or weak realized contribution.

This comparison connects task-conditioned mechanistic analysis with Jacobian-based sensitivity analysis. Existing work characterizes attention heads through attention patterns, output-value circuitry, interventions, and task-specific circuits (Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023; Hanna et al., 2023), while a separate literature uses Jacobians for sensitivity, robustness, intermediate representations, semantic readouts, and token-level attribution (Novak et al., 2018; Jukić & Šnajder, 2025; Gurnee et al., 2026; Liu et al., 2026). To our knowledge, prior work has not systematically compared every head's output-to-vocabulary Jacobian with both task-directed gradients and behavioral readouts across multiple language models.

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

We evaluate GPT-2 Small, Pythia-410M, and GPT-Neo-125M. Indirect object identification (IOI) is the main controlled task; the greater-than task supplies a second output contrast for testing task specificity. Finite interventions test whether large local Jacobian responses remain predictive at nonzero perturbation scales.

Pythia-410M provides the clearest dissociation: its two most broadly responsive heads among all 384 rank only 263rd and 374th in IOI behavioral visibility. Finite interventions nevertheless closely match the Jacobian-predicted slopes, with approximately 1% relative error across the tested scales, while normalized output gain remains comparable to sensitivity-matched high-response controls. Their finite-response alignment increases under the greater-than contrast, but matched controls show a similar greater-than level, so this is evidence of task-dependent geometry rather than greater-than behavioral importance. GPT-2 shows stronger coupling between broad response and behavior, while GPT-Neo supplies supporting high-response/weak-DLA evidence rather than a matched behavioral replication.

2 Related Work

Task-conditioned attention-head and circuit analysis. Attention heads have been studied through pruning, ablation, linguistic specialization, activation patching, and circuit discovery (Michel et al., 2019; Voita et al., 2019; Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023; Conmy et al., 2023). These methods measure causal or additive relevance to a chosen behavior. First-order attribution patching makes this dependence explicit by combining a task gradient with a clean–corrupted activation difference (Syed et al., 2024); our task-directed Jacobian term isolates the gradient-side local sensitivity before that realized activation difference is applied. Benchmarks such as MIB likewise evaluate localization with respect to task performance (Mueller et al., 2025). Such methods remain essential for behavioral attribution, but a low task-conditioned score does not distinguish low downstream gain from high gain that is poorly aligned with, or weakly realized in, the selected behavior. Related work on task-dependent head function and modular circuits further motivates separating task-specific use from broader component characterization (Park et al., 2026; He et al., 2025).

General and output-centric interpretation. Information Flow Routes and ModCirc seek reusable or task-independent structure through attribution graphs and modular circuits (Ferrando & Voita, 2024; He et al., 2025). Output-centric feature interpretation similarly argues that activation-based descriptions can miss how a feature affects model outputs (Gur-Arieh et al., 2025). Our object is different: rather than assigning semantic descriptions to latent features or discovering routes, we quantify the local vocabulary-output response of each attention head and compare it with task-directed and behavioral readouts.

Jacobians and local linearization. Local linearization has been used to study relational extraction, backward vocabulary projections, input-dependent transformer maps, global operator views, and input-to-output attribution (Hernandez et al., 2024; Katz et al., 2024; Golden, 2025; Atad et al., 2026; Liu et al., 2026). The closest output-oriented method is Jacobian Lens, which constructs layer-level vocabulary readouts by averaging residual-stream-to-final-state Jacobians across positions and contexts (Gurnee et al., 2026). We instead compute a prompt-conditioned downstream Jacobian from each individual attention-head output to the full vocabulary logits, aggregate broad and task-directed response, compare both with patching and DLA, and validate selected high-gain regimes by finite intervention. The contribution is therefore not a replacement for task-specific importance, but a complementary measurement of locally available downstream response that task-conditioned rankings alone do not identify.

3 Head-Resolved Output-Jacobian Analysis

Task-conditioned importance and local downstream responsiveness are not equivalent. For a head-level activation change,

$$\Delta \boldsymbol{\ell} \approx \mathbf{J}_h(x)\, \Delta \mathbf{z}_h, \tag{1}$$

so weak behavioral realization can arise from low downstream gain, poor task-direction alignment, a small or poorly oriented natural activation change $\Delta \mathbf{z}_h$, or a combination of these factors. We there-

fore measure three views of a head: broad local output response, task-directed sensitivity/alignment, and behavioral visibility under the selected readouts. We do not attempt to identify the natural modulation mechanism $\Delta \mathbf{z}_h$ separately.

3.1 Local Downstream Response

We first ask how strongly a small change in one head can change the model's vocabulary predictions, without choosing a particular task direction. For an input x, let $\boldsymbol{\ell}(x) \in \mathbb{R}^{|\mathcal{V}|}$ denote the vocabulary logits at the prediction position. For head h, let $\mathbf{z}_h(x) \in \mathbb{R}^{d_h}$ denote its final-position hook_z activation: the value-weighted head output before its head-specific output projection. Treating this activation as an intervention variable while holding the input fixed, we define the head-resolved downstream Jacobian

$$\begin{aligned} \mathbf{J}_h(x) &= \frac{\partial \boldsymbol{\ell}(x)}{\partial \mathbf{z}_h(x)}, \\ \mathbf{J}_h(x) &\in \mathbb{R}^{|\mathcal{V}| \times d_h}. \end{aligned} \tag{2}$$

Differentiation starts at the head output $\mathbf{z}_h$, so the map includes the head-specific output projection and all downstream transformer computation. For a sufficiently small perturbation $\boldsymbol{\delta}$, the first-order output change is $\mathbf{J}_h(x)\boldsymbol{\delta}$, with approximation error $O(\|\boldsymbol{\delta}\|_2^2)$. Intuitively, $\mathbf{J}_h(x)$ tells us how much the model's vocabulary logits are locally able to move when that head is changed.

We do not need to construct the full Jacobian explicitly. Instead, we draw K random unit directions in vocabulary-logit space and estimate

$$\begin{aligned} \widehat{R}_h(x) &= \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{J}_h(x)^\top \mathbf{u}_k \right\|_2, \\ \mathbf{u}_k &= \frac{\widetilde{\mathbf{u}}_k}{\|\widetilde{\mathbf{u}}_k\|_2}, \\ \widetilde{\mathbf{u}}_k &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{3}$$

We evaluate this quantity efficiently with vector–Jacobian products and call it the *broad output-Jacobian response*. It is a direction-averaged response measure, not a spectral norm. "Direction-agnostic" means only that no task-specific output contrast is selected: $\widehat{R}_h(x)$ remains prompt-conditioned and local to the realized final-position state. Head-level scores average $\widehat{R}_h(x)$ across prompts.

3.2 Task Projection and Behavioral Realization

Broad response does not tell us whether the available output change points in the direction that matters for a particular task. We therefore next measure how strongly the same head can affect a selected task decision. Let $\mathbf{d}_x \in \mathbb{R}^{|\mathcal{V}|}$ encode the task-relevant logit contrast. We define the task metric and corresponding head sensitivity as

$$\begin{aligned} m(x) &= \mathbf{d}_x^\top \boldsymbol{\ell}(x), \\ T_h(x) &= \|\nabla_{\mathbf{z}_h} m(x)\|_2 = \left\| \mathbf{J}_h(x)^\top \mathbf{d}_x \right\|_2. \end{aligned} \tag{4}$$

For IOI, $\mathbf{d}_x = \mathbf{e}_{\mathrm{IO}(x)} - \mathbf{e}_{\mathrm{S}(x)}$, so $m(x)$ is the indirect-object-minus-subject logit difference. Greater-than uses the corresponding valid-versus-invalid completion contrast (Section 4). Since T_h is the gradient magnitude of the task metric as defined, its absolute scale inherits $\|\mathbf{d}_x\|_2$; raw values are interpreted within a task, while cross-task comparisons below use normalized finite-response alignment. A head can therefore have high $\widehat{R}_h$ but low T_h when much of its available response lies away from the selected task direction.

Neither derivative measures realized use. The term $\mathbf{J}_h^\top \mathbf{d}_x$ is the task-gradient factor in a first-order attribution-patching approximation; the realized approximation also depends on the clean–corrupted

activation difference $\Delta \mathbf{z}_h$ (Syed et al., 2024). We therefore compare the derivatives with activation patching and direct logit attribution (DLA) (Elhage et al., 2021; Wang et al., 2023). Patching substitutes the clean final-position head activation into a corrupted prompt and measures the change in $m(x)$; DLA measures the head's additive logit contribution to the same task direction in the natural forward pass.

For the canonical Pythia analysis, let $\overline{P}_h$ and $\overline{D}_h$ denote the prompt-aggregated patching and DLA effects. Both are expressed in logit-difference units. We define behavioral visibility conservatively as

$$B_h = \max\Big\{|\overline{P}_h|, |\overline{D}_h|\Big\}. \tag{5}$$

A head will only have a low behavioral rank if neither readout gives it a large task effect. The combined score is then applied in the canonical head-ranking; the results from the patching and DLA remain distinct in the following key diagnostics.

3.3 Finite-Response Validation

A local derivative is useful only if it corresponds to a realizable response at a nonzero perturbation size. We therefore test whether the Jacobian predicts finite changes in model output. For a fresh random unit output direction $\mathbf{u}$, we compute its gradient in the target head space and perturb the clean activation along the normalized gradient:

$$\begin{aligned} \mathbf{g}_h &= \mathbf{J}_h(x)^\top \mathbf{u}, \\ \mathbf{v}_h &= \frac{\mathbf{g}_h}{\|\mathbf{g}_h\|_2}, \\ \mathbf{z}_h^{\pm} &= \mathbf{z}_h \pm \epsilon \|\mathbf{z}_h\|_2 \mathbf{v}_h. \end{aligned} \tag{6}$$

The direction $\mathbf{v}_h$ locally maximizes the selected output projection among unit perturbations in head space. Scaling by $\|\mathbf{z}_h\|_2$ makes ϵ a relative perturbation size.

Let $\boldsymbol{\ell}_h^+$ and $\boldsymbol{\ell}_h^-$ denote the logits produced by the two interventions, and define their symmetric displacement as $\Delta\boldsymbol{\ell}_h = \boldsymbol{\ell}_h^+ - \boldsymbol{\ell}_h^-$. The Jacobian predicts the directional slope $\|\mathbf{g}_h\|_2$; its finite-difference counterpart is

$$\begin{aligned} s_h^{\text{pred}} &= \|\mathbf{g}_h\|_2, \\ s_h^{\text{fin}} &= \frac{\mathbf{u}^\top \Delta\boldsymbol{\ell}_h}{2\epsilon\|\mathbf{z}_h\|_2}. \end{aligned} \tag{7}$$

We compare the slopes by relative error with a small numerical floor. The symmetric central-difference estimator has $O((\epsilon\|\mathbf{z}_h\|_2)^2)$ truncation error under standard smoothness assumptions, distinct from the generic $O(\|\boldsymbol{\delta}\|_2^2)$ remainder of a first-order displacement approximation. Agreement tests whether the local derivative remains predictive at nonzero scale. We also measure normalized finite gain and task-direction alignment:

$$\begin{aligned} G_h &= \frac{\|\Delta\boldsymbol{\ell}_h\|_2}{2\epsilon\|\mathbf{z}_h\|_2}, \\ A_h &= \frac{\left|\widehat{\mathbf{d}}_x^\top \Delta\boldsymbol{\ell}_h\right|}{\|\Delta\boldsymbol{\ell}_h\|_2}, \\ \widehat{\mathbf{d}}_x &= \frac{\mathbf{d}_x}{\|\mathbf{d}_x\|_2}. \end{aligned} \tag{8}$$

Here A_h is the absolute cosine alignment between the finite displacement and the normalized task contrast. Because $\mathbf{v}_h$ is chosen from a fresh random output projection rather than to optimize the task metric, A_h is a geometric diagnostic, not a task-steering intervention or evidence of causal use. For IOI we also directly measure the absolute finite task slope, $|\mathbf{d}_x^\top \Delta\boldsymbol{\ell}_h|/(2\epsilon\|\mathbf{z}_h\|_2)$. At the

head level we report the descriptive product $G_h A_h$ as *alignment-weighted finite gain*; because the component quantities are prompt-aggregated separately, this product is kept distinct from the directly measured finite task slope. Thus $\widehat{R}_h$ and G_h characterize broad local response, T_h and A_h characterize task-directed sensitivity and geometric alignment, and patching/DLA provide behavioral readouts.

4 Experimental Setup

4.1 Models and tasks.

We evaluate GPT-2 Small, Pythia-410M, and GPT-Neo-125M (Radford et al., 2019; Biderman et al., 2023; Black et al., 2021), containing 144, 384, and 144 attention heads. Using TransformerLens (Nanda & Bloom, 2022), we keep pretrained parameters fixed. Jacobian measurements and patching operate on each head's final-position hook_z activation before its output projection; DLA uses the corresponding projected residual-stream contribution after final layer normalization.

Our primary task is indirect object identification (IOI) (Wang et al., 2023). GPT-2 and Pythia use 500 unique IOI prompts sampled from 600 concrete templates and a single-token name pool; the task metric is the indirect-object-minus-subject logit difference from Section 3.2. Selected Pythia heads are additionally evaluated on the greater-than task (Hanna et al., 2023), using the mean logit of valid continuations $YY + 1, \ldots, 99$ minus that of invalid continuations $00, \ldots, YY$. Selection uses IOI only, so greater-than provides a separate task direction. Table 1 summarizes coverage.

Table 1: Experimental coverage. "Screen" and "Finite" report the numbers of IOI prompts used for all-head screening and finite-response validation. "Probes" reports the number of random vocabulary directions evaluated per screening prompt.

Model	Screen	Probes	Finite
GPT-2	500	32	*Not needed*
Pythia	500	16	150
GPT-Neo	150	1	75

4.2 Screening and head selection.

For GPT-2 and Pythia, we screen every head on the 500-prompt IOI set and measure behavioral visibility with activation patching over 25 matched clean–corrupted pairs and DLA over 40 prompts. Each pair uses the same sentence frame but four distinct names: the clean prompt has $(A, B, S = A)$, with B as the indirect object, whereas the corrupted prompt uses $(C, D, S = C)$. Patching is scored on the original clean IO-versus-subject contrast, so neither clean target name appears in the corrupted prompt. Pythia's behavioral score B_h is the maximum absolute patching or DLA effect, ensuring that a low rank requires weakness under both readouts.

The canonical Pythia group contains heads in the top 10% of broad response and bottom 50% of behavioral visibility. Each selected head is paired with a unique behaviorally strong control having the closest broad-response score; a high-behavior group provides a second reference. Selection is independent of task-alignment, finite-response, and cross-task results. We use 2,000 prompt-level bootstrap resamples to test stability of broad-response top-decile membership; the behavioral bottom-half classification is not re-estimated in this bootstrap.

No GPT-2 head passes the same high-response/weak-behavior screen, so GPT-2 is analyzed through all-head and depth-controlled relationships rather than a separately selected finite-validation group. GPT-Neo provides a supporting third-architecture comparison: all 144 heads are screened on 150 prompts with one random vocabulary probe per prompt and behavioral visibility is ranked by mean absolute direct IOI projection. Because GPT-Neo lacks matched activation patching and uses an unnormalized projection, its values are interpreted only within model. We characterize five high-response/weak-DLA heads and controls, but do not treat this as a matched replication of the Pythia behavioral analysis.

4.3 FINITE VALIDATION AND STATISTICS.

Pythia finite interventions use 150 IOI prompts at $\epsilon \in \{0.01, 0.03, 0.10\}$, with $\epsilon = 0.03$ primary, and 150 greater-than prompts at $\epsilon = 0.03$. GPT-Neo uses 75 IOI prompts at the primary scale. For each validation prompt, a fresh random vocabulary direction is shared across evaluated heads; the interventions measure Jacobian prediction error, normalized vocabulary-logit gain, and task-direction alignment.

All-head relationships use Spearman correlation. For GPT-2, depth-controlled relationships are additionally assessed with 2,000 layer-level permutations. Selected-versus-matched comparisons use two-sided paired Wilcoxon signed-rank tests; comparisons with high-behavior controls use two-sided Mann–Whitney U tests. With only five matched pairs, the exact two-sided Wilcoxon test has minimum attainable nonzero $p = 0.0625$ when all differences share one sign, so these tests are treated as low-resolution group diagnostics rather than a standalone significance gate. Reproducibility details are in Appendix A; supplementary statistics and robustness diagnostics are in Appendices B and C.

5 RESULTS

Broad response, task alignment, and behavioral visibility are not interchangeable. Pythia provides the clearest high-response/weak-behavior cases; GPT-2 and GPT-Neo establish that the relationship among the measures is model-dependent. We first compare all heads and then finite-validate the selected Pythia regime.

5.1 TASK-SPECIFIC IMPORTANCE CAN MISS BROADLY RESPONSIVE HEADS

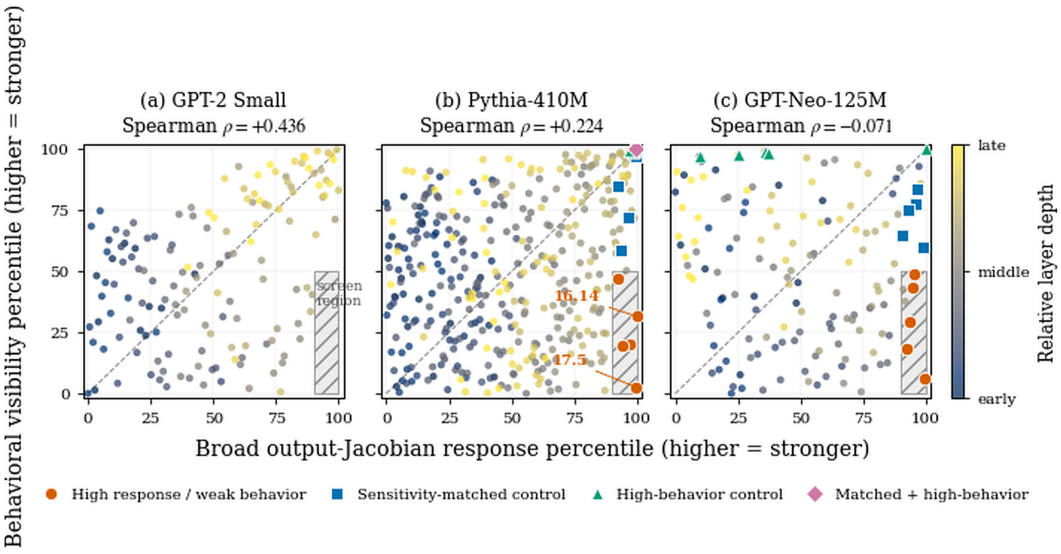

Figure 1: Broad output-Jacobian response versus behavioral visibility. Each point is an attention head; both axes show within-model percentiles, and color denotes relative layer depth. The shaded region marks the top-decile-response and lower-half-behavior screen. Pythia and GPT-Neo contain strongly responsive heads with weak behavioral rankings.

Figure 1 compares all heads under broad output-Jacobian response and behavioral visibility. GPT-2 shows moderate coupling between the rankings ($\rho = 0.436$), and no head satisfies the strict high-response/weak-behavior criterion. This association is partly depth-dependent: the correlation between broad response and absolute activation patching falls from $\rho = 0.706$ globally to a mean of $\rho = 0.258$ within layers. In contrast, the task-directed association remains stable, changing from $\rho = 0.320$ to $\rho = 0.332$ (permutation $p = 0.0005$).

Pythia shows weaker population coupling ($\rho = 0.224$, $p = 9.16 \times 10^{-6}$) and a sharp extreme-rank dissociation: the two most broadly responsive heads rank only 263rd and 374th in behavioral visibility, with three additional top-decile-response heads below the behavioral median. Because B_h takes the larger absolute patching or DLA effect, these heads are weak under both IOI readouts despite high local response.

GPT-Neo provides supporting, but methodologically weaker, evidence: broad response and absolute DLA are uncorrelated ($\rho = -0.071$, $p = 0.401$), and seven of 144 heads satisfy the same rank-based screen. Because its DLA projection is unnormalized and no matched patching readout is used, we interpret this only as a within-model high-response/weak-DLA regime.

5.2 Finite interventions confirm real output response and characterize task alignment

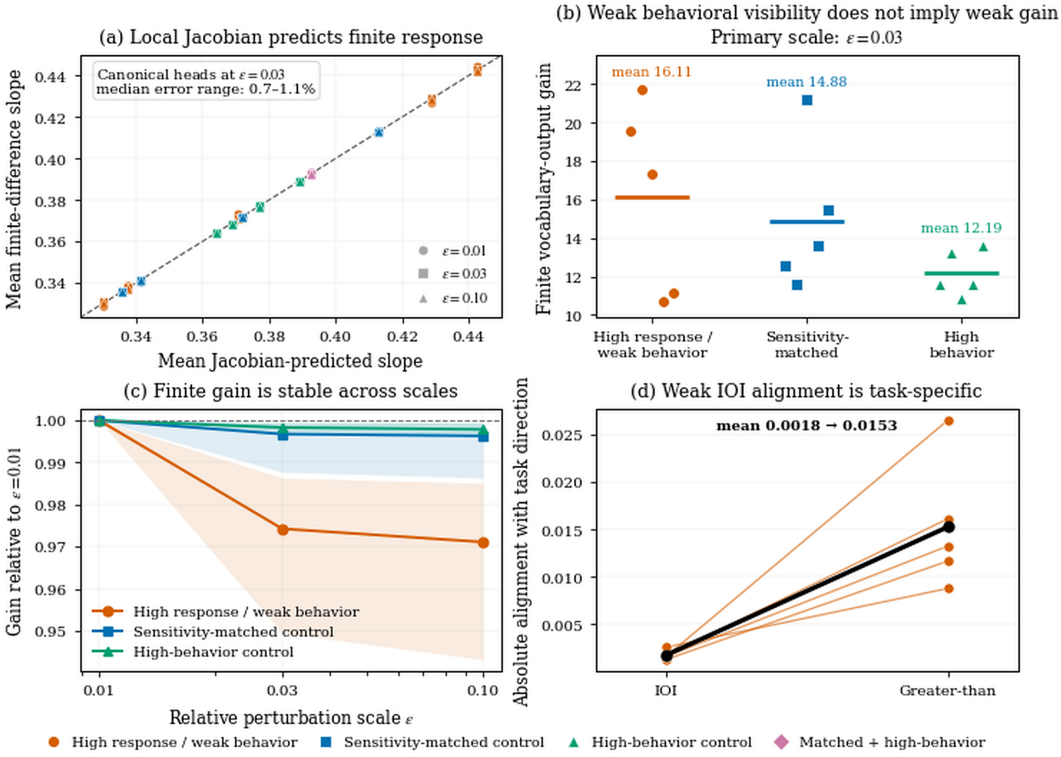

Figure 2: Finite validation and task alignment in Pythia-410M. (a) Jacobian predictions closely match realized finite-difference slopes. (b) Selected heads retain normalized vocabulary-output gain comparable to sensitivity-matched high-response controls. (c) Gain remains stable across perturbation scales. (d) Selected-head alignment increases under greater-than, while matched controls reach a similar greater-than level.

We finite-validate five selected Pythia heads, five sensitivity-matched controls, and five high-behavior controls; one head overlaps the two control groups, giving 14 unique heads. All groups were fixed before finite-response, alignment, and cross-task analysis.

Figure 2(a) shows close agreement between Jacobian-predicted and realized slopes for $\epsilon \in \{0.01, 0.03, 0.10\}$. At the primary scale $\epsilon = 0.03$, median relative errors for the selected heads are 0.7%–1.1%, confirming that the measured local responses correspond to finite output changes rather than numerical gradient artifacts.

The selected heads retain normalized vocabulary-wide gain comparable to the sensitivity-matched high-response controls: mean $G_h = 16.11$ versus 14.88 (paired $p = 0.8125$), and 12.19 for high-

behavior controls (Figure 2(b)). Gain is also roughly stable across the tested perturbation scales (Figure 2(c)). Their weak IOI visibility therefore does not imply weak local downstream responsiveness.

The selected group has lower mean absolute IOI alignment (0.0018) than sensitivity-matched and high-behavior controls (0.0040 and 0.0039). Across the 14 unique heads, finite gain is not associated with behavioral visibility ($\rho = -0.380$, $p = 0.180$), whereas absolute alignment is ($\rho = 0.727$, $p = 0.003$); the directly measured absolute finite IOI slope is also associated with behavior ($\rho = 0.604$, $p = 0.022$). Thus high normalized gain can coexist with little response along the IOI contrast.

Alignment is task-dependent: every selected head increases under the greater-than contrast, with mean A_h rising from 0.0018 on IOI to 0.0153 (Figure 2(d)). Sensitivity-matched controls, however, reach a similar greater-than mean of 0.0144 (paired $p = 1.000$). The cross-task result therefore shows that the selected heads are not uniformly low-alignment channels; it does not establish special greater-than functionality or causal use.

5.3 Robustness and cross-model scope

Appendix C separates broad-response stability from activation scale. All 14 characterized Pythia heads remain in the top decile of broad response in all 2,000 prompt-level bootstrap resamples. Yet the selected heads have much smaller mean final-position activation norm (0.581) than sensitivity-matched and high-behavior controls (2.636 and 3.776). The selected group is therefore not characterized by unusually large activation magnitude. At the same time, these norms bound the interpretation of G_h: because the intervention is scaled by $\|\mathbf{z}_h\|_2$, high normalized gain is not the same as large absolute or naturally realized displacement. Activation norm is also not the task-induced $\Delta\mathbf{z}_h$, so these data do not identify the modulation mechanism behind weak behavioral realization. BOS attention remains descriptive rather than causal.

GPT-Neo independently shows high normalized finite gain among weak-DLA heads (mean 119.20 versus 110.95 and 89.38 for matched and high-behavior controls), but its mean IOI alignment (0.0022) exceeds that of matched controls (0.0014). Pythia's exact low-alignment pattern therefore does not replicate. Across models, the framework reveals different relationships among broad response, task-directed alignment, and behavioral visibility rather than a universal architectural pattern.

6 Conclusion

We introduced a head-level output-Jacobian framework that separates broad local downstream response, task-directed sensitivity/alignment, and behavioral visibility at the final prediction position. Pythia-410M shows why the distinction matters: its two most broadly responsive heads rank only 263rd and 374th in IOI behavioral visibility, yet finite interventions closely match the Jacobian-predicted slopes (approximately 1% relative error) and retain normalized gain comparable to sensitivity-matched high-response controls. Their weak IOI ranking therefore cannot be interpreted as a general absence of local downstream response. At the same time, low IOI alignment and small activation norms show why normalized gain should not be equated with natural behavioral effect or reduced to alignment alone.

GPT-2 and GPT-Neo exhibit different relationships among the same measurements, so the Pythia pattern is not presented as universal. The practical implication is narrower and more general: task-conditioned importance should be interpreted as evidence about the selected behavior, not as a complete measure of a component's local downstream responsiveness.

Limitations

The evidence covers three relatively small autoregressive language models and two controlled token-level tasks. IOI and greater-than provide clean output contrasts but do not span larger or instruction-tuned models, multilingual or long-context settings, or tasks without simple token-level decisions. We therefore do not claim that Pythia's high-response/low-alignment pattern is universal.

Behavioral measurements are not identical across models. GPT-2 and Pythia use activation patching together with DLA, whereas GPT-Neo uses an absolute, unnormalized DLA projection and only one

random vocabulary probe per screening prompt, compared with 16–32 for the primary models. GPT-Neo is therefore supporting high-response/weak-DLA evidence rather than a matched replication. All Jacobian, patching, and DLA claims concern attention-head outputs at the final prediction position and do not characterize sequence-wide head function or causal roles mediated through earlier positions (Geva et al., 2021; Elhage et al., 2022).

Finally, a Jacobian measures local responsiveness, not the natural activation change $\Delta \mathbf{z}_h$ that determines whether this responsiveness is realized. Finite interventions validate selected response directions only over the tested perturbation sizes; they do not establish global linearity or natural usage. The BOS-attention profile is likewise descriptive rather than causal.

AI USE STATEMENT

ChatGPT (OpenAI, 2026) was used as a supplementary tool for literature-search synthesis; brainstorming and refinement of research questions. During the writing process, it was used in limited capacity to provide simpler explanations of previously introduced concepts, assist with the introduction of new theoretical ideas, and suggest basic outlines. AI-assisted drafting and restructuring were concentrated primarily in Abstract and Introduction, where the wording and narrative organization needed to be revised. By default, AI did not substantively rewrite the main technical sections, including methodology, experimental setup, and results.

Lastly, Grammarly(Grammarly, 2026) was used for grammar checking, sentence improvement, redundancy removal, and general readability. Both tools were used only as supplementary writing and learning aids and were not used as substitutes to understand the implementation or produce the final report. The complete written work was reviewed by us before submission.

REPRODUCIBILITY STATEMENT

The experimental setup, model and task definitions, screening criteria, finite-intervention procedure, statistical tests, prompt construction, tokenizer filters, and fixed random seeds are described in Sections 3–5 and Appendix A. Supplementary group-level comparisons and robustness diagnostics are reported in Appendices B and C.

REFERENCES

Ido Andrew Atad, Itamar Zimerman, Shahar Katz, and Lior Wolf. TensorLens: End-to-end transformer analysis via high-order attention tensors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3452–3468. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.156. URL https://aclanthology.org/2026.acl-long.156/.

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

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5297715. Version 1.0.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac

Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17432–17445. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.965. URL https://aclanthology.org/2024.emnlp-main.965/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

James R. Golden. Large language models are locally linear mappings, 2025. URL https://arxiv.org/abs/2505.24293. arXiv:2505.24293.

Grammarly. Grammarly: Ai writing assistance. https://www.grammarly.com/, 2026. Accessed: 2026-09-26.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5757–5778. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.288. URL https://aclanthology.org/2025.acl-long.288/.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html. arXiv:2607.15495.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76033–76060, 2023. doi: 10.52202/075280-3322. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Yinhan He, Wendy Zheng, Yushun Dong, Yaochen Zhu, Chen Chen, and Jundong Li. Towards global-level mechanistic interpretability: A perspective of modular circuits of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22865–22880. PMLR, 2025. URL https://proceedings.mlr.press/v267/he25x.html.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Josip Jukić and Jan Šnajder. From robustness to improved generalization and calibration in pre-trained language models. *Transactions of the Association for Computational Linguistics*, 13:264–280, 2025. doi: 10.1162/tacl_a_00739. URL https://aclanthology.org/2025.tacl-1.13/.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. Backward lens: Projecting language model gradients into the vocabulary space. In *Proceedings of the 2024 Conference on Empirical Methods*

in Natural Language Processing, pp. 2390–2422. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.142. URL https://aclanthology.org/2024.emnlp-main.142/.

Toni J. B. Liu, Baran Zadeoğlu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. Jacobian scopes: Token-level causal attributions in LLMs, 2026. URL https://arxiv.org/abs/2601.16407. arXiv:2601.16407.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pp. 14014–14024, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Neel Nanda and Joseph Bloom. TransformerLens. Software repository, 2022. URL https://github.com/TransformerLensOrg/TransformerLens.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: An empirical study. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJC2SzZCW.

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

OpenAI. Chatgpt. https://chatgpt.com/, 2026. Accessed: 2026-09-26.

Jungwon Park, Jimyeong Kim, Changin Choi, and Wonjong Rhee. Soft head selection for injecting ICL-derived task embeddings. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27161–27214. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1355. URL https://aclanthology.org/2026.findings-acl.1355/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rvi3Wa768B-.

A Reproducibility Details

A.1 IOI prompt construction

We construct 600 concrete IOI templates by crossing the six sentence frames in Table 2 with the lexical sets summarized in Table 3.

Table 2: IOI sentence frames.

No.	Frame
1	`When {A} and {B} went to the {PLACE}, {S} gave a {OBJ} to`
2	`When {A} and {B} went to the {PLACE}, {S} gave the {OBJ} to`
3	`After {A} and {B} left the {PLACE}, {S} handed the {OBJ} to`
4	`While {A} and {B} were at the {PLACE}, {S} passed the {OBJ} to`
5	`Then {A} and {B} arrived at the {PLACE}, and {S} gave the {OBJ} to`
6	`Once {A} and {B} reached the {PLACE}, {S} handed a {OBJ} to`

For every prompt, $S = A$, so B is the indirect object. The prompt therefore contains the name sequence A, B, A, with B as the correct next-name completion. Duplicate prompts are removed until 500 unique prompts have been collected.

For activation patching, each of the 25 pairs samples four distinct names A, B, C, D and one sentence frame. The clean member is formed with $(A, B, S = A)$ and has clean IO/S targets (B, A). The corrupted member uses $(C, D, S = C)$ in the same frame. Patching effects are measured on the original clean contrast $\ell_B - \ell_A$, so neither clean target name occurs in the corrupted prompt.

Table 3: IOI lexical sets and tokenizer filters.

Component	Values or procedure
Objects	*drink, ball, apple, book, note, key, sandwich, letter, pen, cup*
Places	*store, park, shop, office, school, garden, station, market, library, cafe*
Candidate names	*John, Mary, Tom, James, Dan, Sid, Anna, Paul, Kate, Mark, Lucy, Sam, Emma, Jack, Rose, Luke, Peter, Susan, David, Laura, Kevin, Sarah, Brian, Karen, Steve, Nancy, Scott, Helen, Alice, Henry, Grace, Frank, Diana, Simon, Julia, Adam, Ruth, Eric, Clara, Nick*
Name filter	Names are represented with a preceding space and retained only when tokenized as one token with `prepend_bos=False`. All 40 names pass for GPT-2 Small and Pythia-410M.
GPT-Neo filter	Both candidate names must also be accepted by `to_single_token` under the GPT-Neo tokenizer. All 500 generated IOI prompts pass this compatibility test; 150 are used for screening and 75 for finite validation.

A.2 Greater-than prompt construction

Greater-than prompts have the form

```
The NOUN lasted from the year 17YY to the year 17
```

where `NOUN` is sampled from *war, famine, drought, siege, reign, voyage, plague, feud, dynasty,* and *conflict*, and YY is sampled from 02 through 98. The valid output set contains two-digit continuations greater than YY, while the invalid set contains continuations less than or equal to YY. All strings from `00` through `99` are represented by one token under the Pythia tokenizer.

A.3 SAMPLING AND RANDOM SEEDS

Table 4: Experimental sampling and explicitly fixed seeds.

Procedure	Coverage or setting	Seed
GPT-2 IOI prompt construction	500 prompts	0
Pythia IOI prompt construction	500 prompts	0
Activation-patching pairs	25 clean–corrupted pairs	1
Greater-than prompt construction	500 prompts	2
Pythia prompt bootstrap	2,000 resamples of 500 prompts	2026
Pythia finite IOI validation	150 prompts; $\epsilon \in \{0.01, 0.03, 0.10\}$	2026
Pythia greater-than validation	150 prompts; $\epsilon = 0.03$	2027
GPT-Neo all-head screening	150 prompts	2028
GPT-Neo finite validation	75 prompts; $\epsilon = 0.03$	2029
GPT-2 depth permutation test	2,000 layer-level permutations	0

The original GPT-2 and Pythia all-head random-probe computations do not set an additional explicit PyTorch seed beyond the dataset-construction settings above. Table 4 summarizes the sampling settings and explicitly fixed random seeds used in each procedure.

B SUPPLEMENTARY RESULTS

B.1 PYTHIA HEAD GROUPS

The five high-response/weak-behavior heads are selected using broad response and behavioral visibility only. Task alignment and finite-response measurements are not used during selection. The resulting selected and control groups are listed in Table 5.

Table 5: Canonical Pythia groups.

Group	Heads
High response / weak behavior	16.14, 17.5, 16.2, 19.14, 16.12
Sensitivity-matched	20.15, 19.15, 15.0, 16.11, 15.2
High behavior	19.15, 11.4, 17.6, 16.0, 18.8

Head 19.15 belongs to both control groups, resulting in 14 unique characterized heads.

B.2 GROUP-LEVEL DIAGNOSTICS

Table 6: Group means for Pythia diagnostics.

Group	G_h	A_h	$\|z_h\|_2$	BOS
Selected	16.1068	0.0018	0.581	0.878
Matched	14.8822	0.0040	2.636	0.406
High behavior	12.1888	0.0039	3.776	0.381

Group-level diagnostic means are reported in Table 6. Here, G_h is normalized finite vocabulary-output gain at $\epsilon = 0.03$, A_h is absolute finite-response IOI alignment, $\|z_h\|_2$ is the final-position activation norm, and BOS is the mean attention assigned to the beginning token. Every characterized head remains in the top decile of broad response across all 2,000 prompt-level bootstrap resamples. This bootstrap statement concerns broad-response membership only; the behavioral bottom-half screen is not re-estimated in these resamples. The much smaller activation norm of the selected group should be read together with the normalization in G_h: it shows high normalized gain per relative perturbation, not necessarily high absolute or naturally realized logit displacement. For the greater-than contrast, the selected heads have mean finite-response alignment 0.0153 versus 0.0144

for sensitivity-matched controls (paired two-sided Wilcoxon $p = 1.000$); the corresponding mean normalized finite gains are 19.0360 and 19.5964 ($p = 0.625$). These cross-task diagnostics support a task-specific change in alignment without establishing special greater-than behavioral relevance.

B.3 STATISTICAL COMPARISONS

Matched comparisons use two-sided paired Wilcoxon signed-rank tests. Comparisons with the high-behavior group use two-sided Mann–Whitney U tests. The corresponding group-comparison results are reported in Table 7. With five matched pairs, 0.0625 is the smallest attainable nonzero exact two-sided Wilcoxon p-value when all pairwise differences share one sign; these matched-group tests are therefore interpreted as low-resolution diagnostics rather than decisive significance tests.

Table 7: Pythia group-comparison results at $\epsilon = 0.03$.

Metric	**Matched** p	**High-beh.** p
Normalized finite output gain	0.8125	0.5476
Absolute IOI alignment	0.1250	0.0079
Absolute IOI slope	0.1250	0.0159
Alignment-weighted finite gain $G_h A_h$	0.0625	0.0159

C Robustness and Cross-Model Diagnostics

Figure 3 summarizes bootstrap stability, activation magnitude, attention profiles, GPT-2 depth controls, bidirectional rank disagreements in Pythia, and the supporting GPT-Neo comparison.

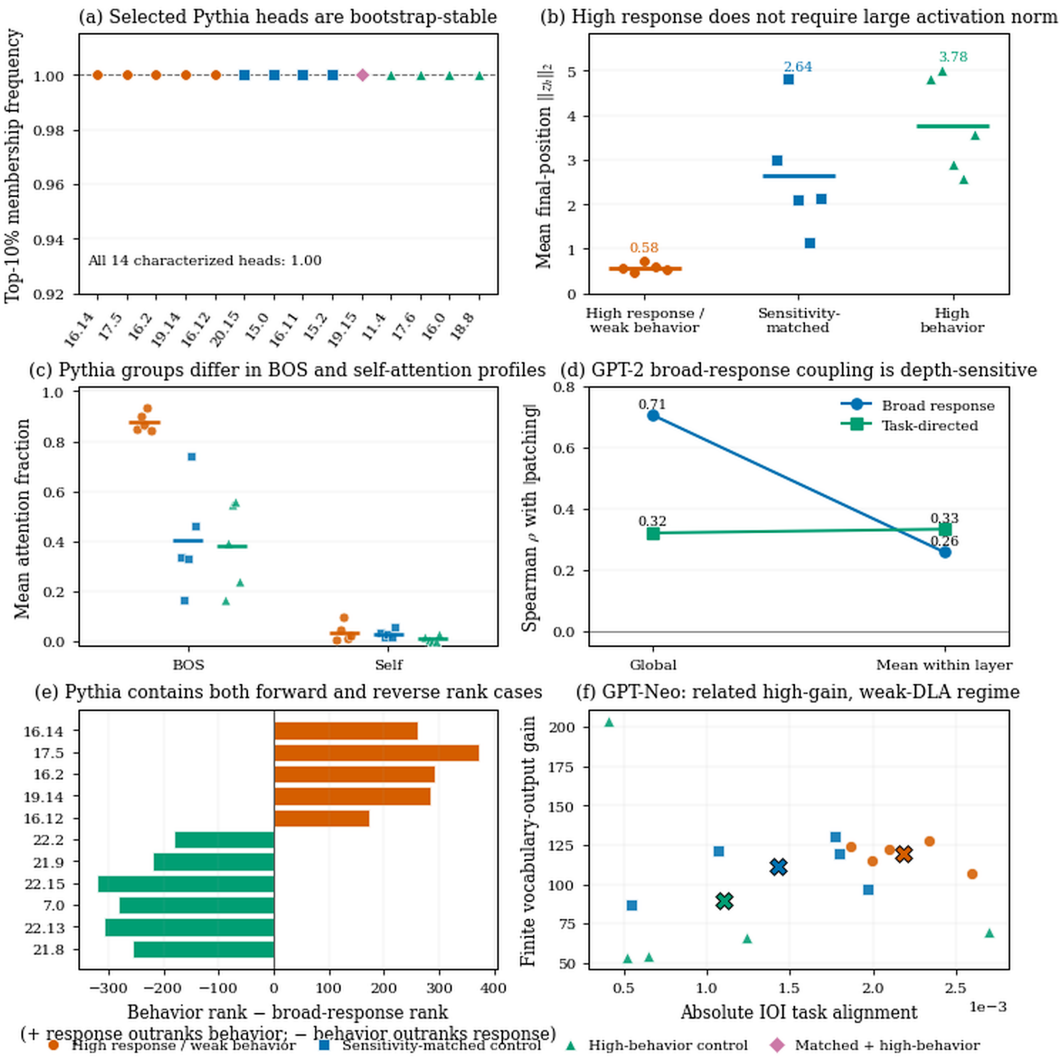

Figure 3: Robustness and cross-model diagnostics. (a) Top-decile membership across 2,000 prompt-level bootstrap resamples for the 14 characterized Pythia heads. (b) Final-position activation norms for selected and control heads. (c) Mean BOS- and self-attention profiles. (d) Global and mean within-layer GPT-2 associations. (e) Response-dominant and behavior-dominant rank disagreements in Pythia. (f) Normalized finite vocabulary-output gain and IOI alignment in GPT-Neo. GPT-Neo behavioral visibility uses absolute, unnormalized DLA and is interpreted only through within-model rankings.