SIMILAR ACCURACY, DIFFERENT MECHANISMS: TRAINING HISTORY, SCALE, AND TASK FORMULATION IN TRANSFORMERS FOR ALZHEIMER'S DISEASE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Similar predictive performance can conceal different computations and clinical decision rules. We study how training history, scale, and task formulation shape decoder-only transformers for ADNI-derived cognitive-status classification. We compare pretrained-and-LoRA-adapted Pythia models from 14M to 1.4B parameters with from-scratch GPT-NeoX controls, including architecture-matched configurations. Both training regimes are evaluated using classifier-head and next-token formulations. Circuit extraction, probing, activation patching, and clinical counterfactuals distinguish represented information from evidence that influences predictions. The 12-layer from-scratch classifier reaches 90% circuit faithfulness using $(7 \pm 5)\%$ of its attention heads, compared with $(70 \pm 21)\%$ for the architecture-matched Pythia-160M classifier; their corresponding macro-F1 values are approximately 0.82 and 0.70. Both estimates use the same mean-ablation procedure and two-point normalisation, with only valid extractions included. Readout patching places peak attention-mediated influence later in the pretrained-and-adapted members of the matched pairs. Linear probes select the same best-encoded cognitive feature in four of five domains across all controls and seeds, yet interventions reveal greater prediction sensitivity to verbal Functional Activities Questionnaire (FAQ) descriptors than to numerical-score edits in most evaluated from-scratch runs, with a seed-dependent exception. Length-matched clinically neutral filler produces few flips in most runs, but up to 59% in the Pythia-1.4B generative formulation. Score-preserving wording changes flip up to 100% of eligible, initially correct cognitively normal predictions in several models, revealing wording sensitivity that persists in the larger pretrained transformers evaluated here. Leave-one-out necessity analyses identify recurring influential FAQ items, including remembering appointments, assembling tax records, and writing checks, under the tested intervention contexts. Models with similar aggregate performance therefore exhibit different internal intervention profiles while showing recurring reliance on functional descriptors.

1 INTRODUCTION

Predictive performance is the dominant criterion for comparing models adapted to clinical tasks. However, two models with similar in-distribution (IID) performance can arrive at the same diagnosis through different internal computations, different clinical variables, or different shortcuts. This distinction matters in high-stakes settings because a model can be correct for a reason that is clinically inappropriate or unstable under small, clinically irrelevant changes to the input. Understanding not only *what* a model predicts, but also *which information causally drives the prediction and how that information is transformed through the network*, is therefore important for scientific understanding and for auditing clinical AI systems (Feng et al., 2023; Reg, 2024).

Our central question is whether **predictive equivalence implies mechanistic equivalence**. We study three key factors; **training history, model scale, and task formulation**. By training history we mean both what a model inherits from language-model pretraining and how it is subsequently

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.

adapted to the clinical task. Recent work shows that fine-tuning can reorganise task-relevant circuits and that some mechanisms can remain stable across training and scale (Wang et al., 2025; Tigges et al., 2024). Separately, modern text classification studies show that label-space and vocabulary-space prediction can achieve comparable headline performance (Kasa et al., 2025; Lyu et al., 2026). These results leave open a different question: when closely matched autoregressive transformers solve the same task through a classification head or through next-token label generation, do they use the same internal computation and the same evidence?

Pretraining complicates this comparison. Downstream adaptation can reuse representations acquired during language modelling, but can also redirect or suppress pretrained behaviour (Kotha et al., 2024). Large language models may also exploit spurious correlations or shortcut features that support strong in-distribution performance (Yuan et al., 2024). Studying only pretrained models therefore makes it difficult to separate inherited representations from mechanisms introduced during downstream adaptation, whereas from-scratch models offer greater experimental control but leave open whether the resulting mechanisms extend to pretrained systems.

We address this tension within a single architectural family under two training regimes. Pythia (Biderman et al., 2023) is a suite of decoder-only transformers built with the GPT-NeoX framework. We adapt models from 14M to 1.4B parameters using either a task-specific classifier head or next-token diagnostic generation. For controlled comparison, we also train GPT-NeoX models from random initialisation with principal dimensions matched to selected Pythia checkpoints: 6 layers, $d = 256$, and 8 heads for Pythia-31M, and 12 layers, $d = 768$, and 12 heads for Pythia-160M, together with a depth-control configuration. We refer to these as *architecture-matched from-scratch controls*. Their purpose is not to serve as substitutes for foundation models, but to reduce architectural confounding and enable tractable circuit- and feature-level interventions.

Crucially, this comparison should not be read as varying pretraining alone. The Pythia models combine large-scale language-model pretraining with LoRA task adaptation, whereas the controls are trained directly on the diagnostic task from random initialization. The matched pairs therefore contrast *training regimes* while holding the model family and major architectural dimensions fixed. This distinction lets us ask which mechanistic differences persist under architectural matching without over-attributing them uniquely to pretraining.

Alzheimer's disease diagnosis from Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Petersen et al., 2010) provides a useful testbed because the input contains individually manipulable demographic, neuropsychological, and functional variables, represented through both numerical scores and verbal descriptions. This lets us move from generic interpretability to a clinically grounded causal question: does a model rely on the intended clinical construct, or on a demographic correlate or surface-form cue? Recent work on Alzheimer's LLM interpretability has similarly highlighted the value of representation- and feature-level explanations (Mamalakis et al., 2026); here we ask whether those features are merely decodable or attributionally salient, or whether intervening on them actually changes the model's decision. All analyses use as input the tokenized representation of the clinical narratives described in Appendix A. The preprocessing and extraction of structured clinical variables into text follow the general principles established in the ADNI clinical-language modelling pipeline of Mamalakis et al. (2026).

Existing work has examined mechanistic changes across fine-tuning and scale, discriminative versus generative prediction, and feature-level explanations in clinical language models (Tigges et al., 2024; Wang et al., 2025; Kasa et al., 2025; Lyu et al., 2026; Mamalakis et al., 2026). We bring these threads together by asking whether models with similar predictive performance also rely on the same internal computation and the same clinical evidence. Extended related work is provided in Appendix A.

Contributions. Our results support three main conclusions. First, predictive performance varies non-monotonically across the adapted Pythia scale ladder: the from-scratch controls achieve similar macro-F1 across configurations, whereas pretrained-and-LoRA-adapted Pythia models vary more across scales and task formulations, with learning-rate tuning reducing some intermediate-scale failures. Second, architecture-matched models can exhibit different internal intervention profiles, including differences in attention-head compressibility, the depth of peak attention-mediated influence, and feature sensitivity despite comparable predictive performance. Third, functional-feature reliance recurs across training regimes, but the sufficiency, context-specific necessity, and wording sensitivity of these features vary across configurations. In particular, score-preserving wording

changes can alter predictions even in the larger pretrained models, showing that reliance on clinically relevant inputs does not guarantee invariance to their verbal presentation.

Together, the experiments test a broader claim: *similar predictive performance under IID evaluation does not imply mechanistic interchangeability*. Across the evaluated training history, scales, and output formulations, similar aggregate IID performance coexists with differences in internal intervention profiles and dependence on numerical and verbal evidence.

2 Experimental Setup

Dataset. We study binary Alzheimer's disease (AD) versus cognitively normal (CN) classification using a balanced ADNI baseline-visit dataset with 1,142 training and 1,142 test examples (Petersen et al., 2010). Each participant is represented by a template-generated clinical summary containing demographic and vital information, psychiatric screening, and six structured neuropsychological and one functional assessment (Appendix B.1). The representation deliberately retains both numerical measurements and their verbal clinical descriptors so that the two can be intervened on separately.

Architecture family and pretrained scale ladder. Pythia is a decoder-only transformer suite trained using the GPT-NeoX framework (Biderman et al., 2023). We use six Pythia scales from 14M to 1.4B parameters. Because the suite shares a common pretraining framework and training recipe across scales, it provides a comparatively controlled setting for studying scale within the pretrained regime (Tigges et al., 2024). For the clinical task, all Pythia checkpoints are adapted with LoRA (Hu et al., 2022) (Appendix B.3, Table 3).

Task formulations. For every Pythia scale we train two variants. In the **classifier** formulation, a two-class linear head maps the final read-out representation to AD or CN. In the **generative** formulation, the causal language-model head is retained and the model is fine-tuned to emit `AD` or `CN` after a `Diagnosis:` cue; prediction is obtained from the two label-token logits at the diagnostic read-out position. Thus, both variants use an autoregressive transformer backbone and the same clinical task, while the output space and training objective differ.

Architecture-matched from-scratch controls. We instantiate GPT-NeoX transformers from random initialization so that the comparison remains within the same decoder-only architectural family. The 6L/$d = 256$/$h = 8$ and 12L/$d = 768$/$h = 12$ configurations match the principal architectural dimensions, namely the number of layers, hidden width, and attention-head count, of Pythia-31M and Pythia-160M, respectively; a 12L/$d = 256$/$h = 8$ configuration provides a depth control. We apply the same classifier and generative task formulations. We refer to these models as *architecture-matched from-scratch controls*. Unless otherwise stated, "matched" refers to these principal dimensions rather than a claim that every implementation or training hyperparameter is identical.

Interpretation of the matched comparison. The matched pairs reduce architectural confounding, but they do not isolate a pure causal effect of pretraining. The Pythia branch is first pretrained as a language model and then adapted with LoRA, whereas the from-scratch branch is optimised directly for the downstream task. Differences between a matched pair may therefore reflect pretrained representations, the adaptation scheme, or an interaction between them. We consequently describe these experiments as a comparison of *training regimes* or *training histories*, and reserve causal claims about pretraining itself for designs that also match the adaptation procedure.

Evaluation logic. We report macro F1 because it exposes asymmetric class failures that can be obscured by accuracy or binary F1. We then combine correlational analyses (logit attribution and linear probing) with causal interventions (mean ablation, activation patching, activation swapping, and input counterfactuals) to distinguish information that is present in a representation from information that the model actually uses.

Evaluation and comparison protocol. We assess model similarity at three levels: adaptation to the diagnostic task, the internal organisation of the decision, and the clinical evidence that changes the output. Here, similarity refers to comparable aggregate IID performance rather than identical predictions for every participant. The architecture-matched comparisons pair the 6L/$d = 256$ control

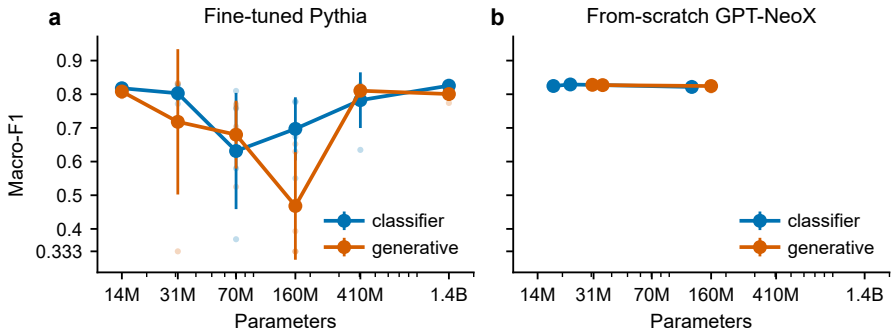


Figure 1: Macro-F1 across training history, task formulation and scale

with Pythia-31M and the 12L/$d = 768$ control with Pythia-160M; the 12L/$d = 256$ model serves as a deep-narrow control. Throughout, these comparisons contrast complete training regimes and task formulations, as defined in Section 1, rather than isolating pretraining or the output head alone. We retain the AD/CN output-label notation used in the experiments.[1] Full training, mechanistic, and intervention protocols are provided in Appendices B, C, and D, respectively.

3 Results

3.1 Predictive similarity coexists with training-regime-dependent adaptation

The scale ladder is non-monotonic even after tuning. The from-scratch controls achieve mean macro-F1 of approximately 0.82–0.83 across the evaluated depths, widths, and task formulations (Figure 1). Comparable performance is also observed at both ends of the pretrained scale ladder: Pythia-14M reaches $0.82/0.81$ macro-F1 for classification/generation, while Pythia-1.4B reaches $0.83/0.80$. Performance is less stable at intermediate scales. At 160M, the classifier achieves 0.70 ± 0.08 macro-F1, whereas the generative formulation reaches 0.47 ± 0.14, with sensitivity 0.76 but specificity 0.35 (Appendix B.4, Table 5). Thus, increasing model size does not consistently improve either discrimination or class balance under the reported adaptation procedure.

Optimisation explains part, but not all, of the intermediate dip. Validation-based learning-rate selection substantially improves several intermediate Pythia models, raising macro-F1 from 0.604 to 0.803 for Pythia-31M classification, from 0.430 to 0.680 for Pythia-70M generation, and from 0.333 to 0.468 for Pythia-160M generation (Appendix B.4, Table 6). The 160M generative model nevertheless remains substantially weaker than its matched from-scratch control. These gains indicate that the initial intermediate-scale failures are partly optimisation-dependent rather than evidence of an intrinsic scaling law, consistent with prior work on fine-tuning stability (Mosbach et al., 2021).

Neither task formulation dominates across the scale ladder. Generation achieves higher macro-F1 at 70M and 410M, whereas classification performs better at 31M, 160M, and 1.4B. At 70M, for example, the higher generative macro-F1 (0.68 versus 0.63) comes with lower sensitivity (0.49 versus 0.79) and higher specificity (0.93 versus 0.55) (Appendix B.4, Table 5). Similar aggregate performance therefore need not imply similar class-wise error profiles. These results identify which models occupy comparable performance regimes and which matched comparisons instead involve differently optimised predictors.

3.2 Training regime and task formulation reorganise diagnostic computation

Matched architectures differ in attention-circuit compressibility. We quantify attention-head compressibility as the fraction of heads required to recover 90% two-point faithfulness relative to the intact and all-head-mean-ablated baselines (Appendix C.1.1, Table 9). Embeddings, MLPs, and

[1] AD here denotes the positive model label. Appendix B.2 notes that the positive cohort is not exclusively Alzheimer's disease.

the task readout remain intact, so this measure captures only the compressibility of the attention computation rather than the fraction of total model computation retained.

The largest architecture-matched classifier contrast appears at 12-layer models (12L): the from-scratch model requires only $(7 \pm 5)\%$ of its 144 attention heads, compared with $(70 \pm 21)\%$ for Pythia-160M (Figure 2). The 6-layer models (6L) differences are smaller: $(51 \pm 16)\%$ versus $(65 \pm 21)\%$ for classification, and $(38 \pm 15)\%$ versus $(51 \pm 23)\%$ for generation. Thus, matching the principal architecture and total head budget does not determine how concentrated the extracted attention circuit will be.

The 12-layer comparison is architecture-matched rather than performance-matched: the corresponding classifier macro-F1 values are approximately 0.82 and 0.70. In contrast, within the six-layer from-scratch configuration, the classifier and generative models both achieve a macro-F1 of 0.83, despite requiring different fractions of attention heads to reach the faithfulness threshold. This relationship is not consistent at larger scale. For Pythia-1.4B, the retained-head fraction ranges from 38.3–97.6% for classification and 53.1–62.8% for generation, with the ordering reversing across the analysed checkpoints.

Circuit sizes are reported only for runs that pass the reliability gates (Appendix C.1); failed extractions are treated as undefined rather than as zero-sized circuits. We therefore interpret the extracted circuits as sufficient subsets under the specified ablation procedure, rather than as globally minimal circuits or complete causal pathways, consistent with prior work on circuit-faithfulness metrics (Miller et al., 2024).

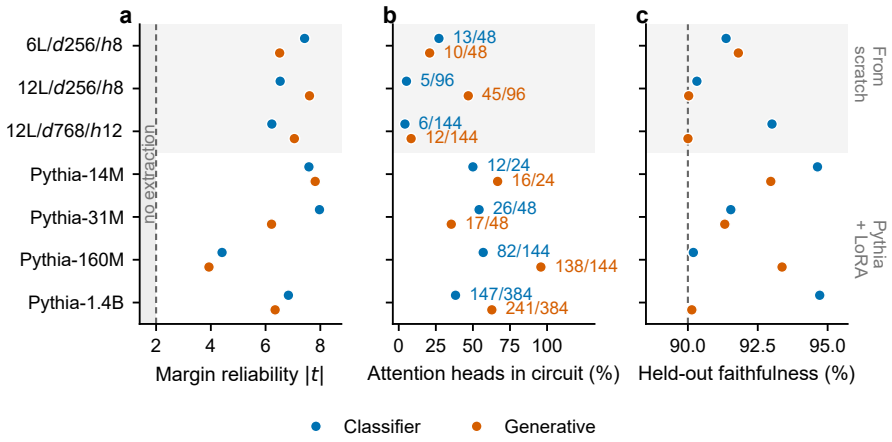


Figure 2: Fraction of attention heads for 90% Faithfulness

The peak of attention-mediated influence shifts across training regimes. Layer-wise readout patching identifies where donor activations exert the strongest effect on the recipient's diagnostic margin (Appendix D.1). In the 6-layer generative pair, attention-output effects peak at layer 0 in the from-scratch controls (median recovery 0.60–0.78), but at layer 4 in Pythia-31M (0.85–0.97). The classifier shows a similar shift, from layer 1 in the from-scratch control to layers 4–5 in Pythia-31M. At 12-layers, attention-output effects again peak at layer 0 in the from-scratch controls, while Pythia-160M peaks at layers 1, 5 or 7 for classification and layers 9 or 11 for generation (Figure 3). Across both donor directions, the peak attention-output effect occurs later in the pretrained-and-adapted member of every reported architecture-matched pair.

This pattern is specific to attention-mediated influence rather than a general delay in computation. The Pythia-160M classifier, for example, shows a strong early MLP patching effect at layer 2 (0.89–1.08)(Appendix D.1.1, Table 15). At 1.4B, attention effects peak around layers 10–11 of 24, indicating that relative peak depth does not increase monotonically with parameter count. At this scale, generation also shows larger peak attention effects (0.46–0.70) than classification (0.24–0.40) (Appendix D.1.1, Table 15). These estimates are conditional on the task-specific readout and model-specific eligible pairs, and the peak should not be interpreted as the first layer at which diagnostic information appears (Zhang & Nanda, 2024).

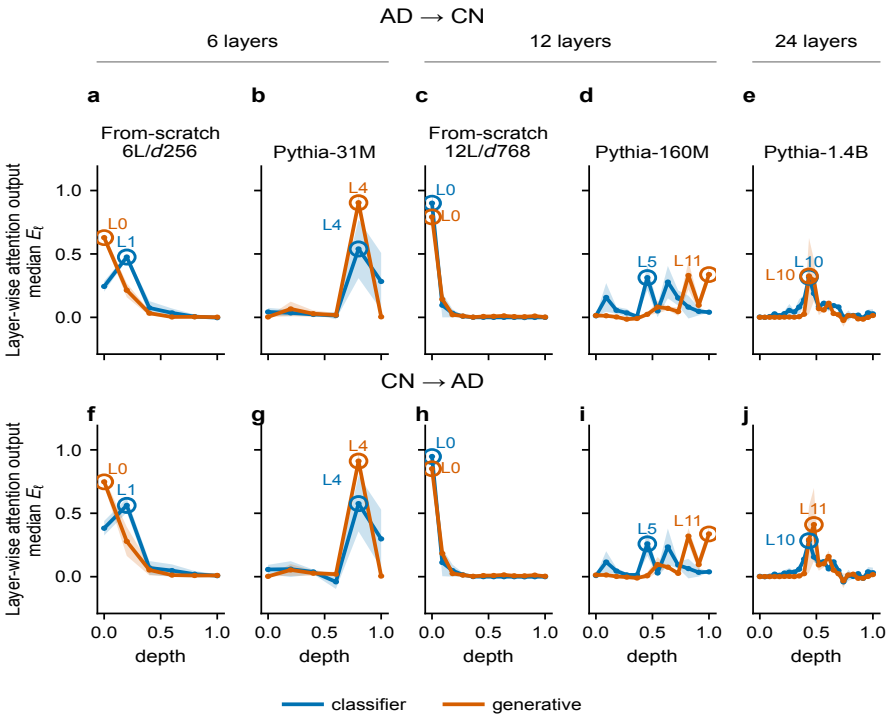


Figure 3: Layer-wise readout activation patching using exactly matched YOB-sex pairs with opposite labels, n=403. Full results available in Appendix D.1.1, Table 15

Descriptive attribution corroborates different organisations, not a shared path. The strongest attributed attention heads are concentrated in layers 0–1 of the 6-layer and wide 12-layer from-scratch controls, but shift toward middle and later layers in Pythia (Appendix C.2.1, Table 10). MLP alignment also differs across formulations (Appendix C.2.1, Table 11). In the 6-layer controls, classifier MLPs are fully CN-directed ($C = -1$), whereas the generative models show near-complete cancellation ($C \in [-0.06, 0.05]$). At 1.4B, this pattern reverses: generation shows more concordant AD-directed attribution ($C = 0.48$–0.55) than classification ($C = 0.01$–0.24), where C is the net MLP attribution divided by its gross magnitude. These quantities measure component-output alignment rather than causal contribution. Together with circuit extraction and patching, they reveal differences in the concentration, depth, and direction of computation across formulations.

3.3 Functional evidence recurs, but its causal role and score invariance differ

Shared decodability does not imply shared use. Verbal inputs are more influential than numerical ones. The from-scratch models strongly encode numerical functional information: FAQ Total reaches a peak held-out R^2 of 0.856–0.980, while the appointments-related FAQ variable reaches 0.743–0.836 (Appendix C.4, Table 14). However, decodability does not imply causal use. Under numerical-token activation swapping, changing the FAQ Total representation produces no prediction flips in most evaluated from-scratch runs (Appendix D.3.1, Table 18). The wide 12-layer generative control is a notable exception, with sharply different behaviour across the two analysed solutions. Although both strongly encode FAQ Total ($R^2 = 0.885$ and 0.856), one produces 99% CN→AD and 17% AD→CN flips under FAQ Total activation swapping, whereas the other produces none (Appendix D.3.1, Table 18). Input-level counterfactuals show the same contrast (Table 1): in the first solution, changing FAQ Total alone flips 100% of eligible CN predictions, compared with 26% for verbal FAQ descriptors; in the second, the corresponding rates are 0% and 100% (Appendix D.4.1, Table 19). Overall, most from-scratch runs are more sensitive to verbal FAQ descriptors than to numerical representations, but the wide 12-layer generative control shows that otherwise similar solutions can rely on different input channels.

Table 1: Fraction of predictions flipped to AD by each input edit for the classifier and generative formulations. Results are shown for two training seeds (s0, s2), with $n = 266$–317 narratives per seed and formulation.

	FAQ total only		FAQ words only		Same-score wording		Matched filler	
Model	s0	s2	s0	s2	s0	s2	s0	s2
Classifier								
6L/*d*256 scratch	0.03	0.01	1.00	1.00	0.92	1.00	0.00	0.00
12L/*d*256 scratch	0.00	0.00	1.00	1.00	0.01	1.00	0.00	0.00
12L/*d*768 scratch	0.00	0.00	0.98	1.00	0.40	0.38	0.01	0.00
Pythia-1.4B pretrained	0.95	0.02	0.33	1.00	0.89	1.00	0.00	0.01
Generative								
6L/*d*256 scratch	0.00	0.08	1.00	1.00	0.68	1.00	0.00	0.00
12L/*d*256 scratch	0.00	0.06	1.00	1.00	0.25	1.00	0.00	0.00
12L/*d*768 scratch	1.00	0.00	0.26	1.00	0.02	0.63	0.01	0.00
Pythia-1.4B pretrained	0.00	0.96	0.88	0.84	0.73	0.99	0.59	0.00

Larger scale and pretraining do not guarantee clinical faithfulness. We next ask whether sensitivity to verbal FAQ descriptors reflects clinical content, surface wording, or sequence length (Table 1). Replacing `Normal (0)` with the score-preserving phrase `Never did, but could do now (0)` flips 89–100% of eligible CN→AD predictions in the Pythia-1.4B classifiers and 38–40% in the wide 12-layer from-scratch classifiers. One possible explanation is lexical overlap with the higher-severity descriptor `Never did, would have difficulty now (1)`. These interventions therefore establish sensitivity to score-preserving wording changes, but do not identify the linguistic mechanism responsible.

Length alone does not explain this effect in most models: matched clinically neutral filler produces only 0–1% flips in nearly all evaluated runs. The main exception is one Pythia-1.4B generative solution, where filler induces up to 59% CN→AD flips (Appendix D.4.1, Table 19). Thus, greater scale and pretrained adaptation do not by themselves guarantee invariance to clinically irrelevant surface-form perturbations.

Demographic activation swaps provide a complementary control. Swapping representations associated with sex, race, or year of birth in the from-scratch models produces at most 4.5% label flips, with sex and race effects near zero in most runs (Appendix D.2.1, Table 16, Table 17). These small effects contrast with the substantially larger responses to verbal functional descriptors.

Joint edits can change predictions when individual edits have weak effects. Given the stronger response to verbal features in most evaluated runs, we next ask which verbal clinical features are sufficient or necessary to change the model's prediction (Figure 4; Appendix D.5.1, Table 20). We summarise individual-item sufficiency using $K_{90} = \sum_{j \in \mathrm{FAQ}} \mathbf{1}\{\mathrm{Suff}_j \geq 0.90\}$, the number of FAQ items whose impairment-inducing edit flips at least 90% of eligible CN predictions. We focus on remembering appointments, assembling tax records, and writing checks because these items recur in the normalising interventions, alongside a control that rewrites all verbal features.

Individual-item sensitivity becomes weaker in several larger configurations (Figure 4d). In particular, the wide 12-layer generative and Pythia-1.4B models have low K_{90} and correspondingly low item-wise necessity for many features (Figure 4a,c). Yet weak single-item effects do not imply weak joint effects: in conditions where no individual FAQ item reaches 50% sufficiency, jointly rewriting all 20 verbal items can still produce 90%–100% flip rates (Figure 4d). For the Pythia-1.4B classifier, rewriting all ten FAQ items flips 55% of eligible CN predictions (Appendix D.5.1, Table 20). These results distinguish weak single-feature sensitivity from strong responses to joint edits, but do not establish nonlinear synergy; several small effects may simply accumulate across a decision boundary, and the single-item and joint rates use different conditioning.

Task formulation also has no consistent ordering. At 6-layers, classifiers have $K_{90} = 0$–3 compared with 8 for generation, whereas at 12-layers the ordering reverses, with $K_{90} = 10$ for classification and 1 for generation, despite comparable macro-F1 (~ 0.83; Figure 1). The same reversal appears

at a 50% sufficiency threshold. Thus, the effect of task formulation on single-feature sensitivity depends on architectural depth rather than following a uniform classifier–generator pattern.

Necessity separates features when sufficiency saturates. Necessity reveals distinctions that are hidden when single-item sufficiency saturates (Figure 4e). In the 12L/d256 deep-narrow classifier, all FAQ items are individually 100% sufficient, yet their necessity differs substantially: withholding the appointments rewrite reduces the AD→CN flip rate by 70–72%, compared with 69% for taxes, 48–57% for checks, and only 13% for heating water (Appendix D.5.1, Table 20). The full all-20 rewrite reaches a 99–100% flip-rate ceiling. These values are context-specific losses, $\mathrm{Nec}_j = F_{\mathrm{all}} - F_{-j}$, rather than unconditional feature importance scores. This distinction complements prior necessity–sufficiency approaches to feature explanation and ADNI interpretation (Chen et al., 2024; Vlontzou et al., 2025), although those studies use different estimators.

Greedy sufficient edit sets differ across models with similar IID performance. We use greedy search to construct sets of jointly edited features that reach a 90% target flip rate, stopping when the threshold is met or all 20 features have been considered. The resulting sets are sufficient under the tested intervention and cohort, but are not guaranteed to be globally minimal (Appendix D.6.1, Table 21).

The results show a pronounced directional asymmetry. In the AD→CN normalising direction, successful sets generally require many edits: the 6-layer generative controls, for example, require 10–11 features to reach a 92% flip rate. In the opposite CN→AD direction, a single impairment-inducing feature is sufficient to flip 100% of eligible predictions in these same models. More broadly, smaller 6L and 12L narrow models often reach the CN→AD threshold with one feature, whereas some 12L wide and 1.4B configurations require several. Thus, the amount of evidence needed to change a prediction depends strongly on both intervention direction and model configuration.

We next ask whether these greedy solutions depend on the identities of the selected features. In the normalising direction, size-matched random subsets recover significantly less of the effect, indicating that cardinality alone does not explain the result (Appendix D.6.1, Table 22). Moreover, in the significant cases reported in Table 23 (Appendix D.6.1), the selected set outperforms its strongest tested one-item substitution. This supports specificity relative to the tested alternatives, rather than global uniqueness or a required sequence of evidence. Remembering appointments, assembling tax records, and writing checks recur among these normalising sets, consistent with prior clinical work identifying related functional activities as discriminative or prognostically informative (Marshall et al., 2015). That literature does not, however, define a ground-truth feature ranking or imply that any single activity is sufficient for diagnosis.

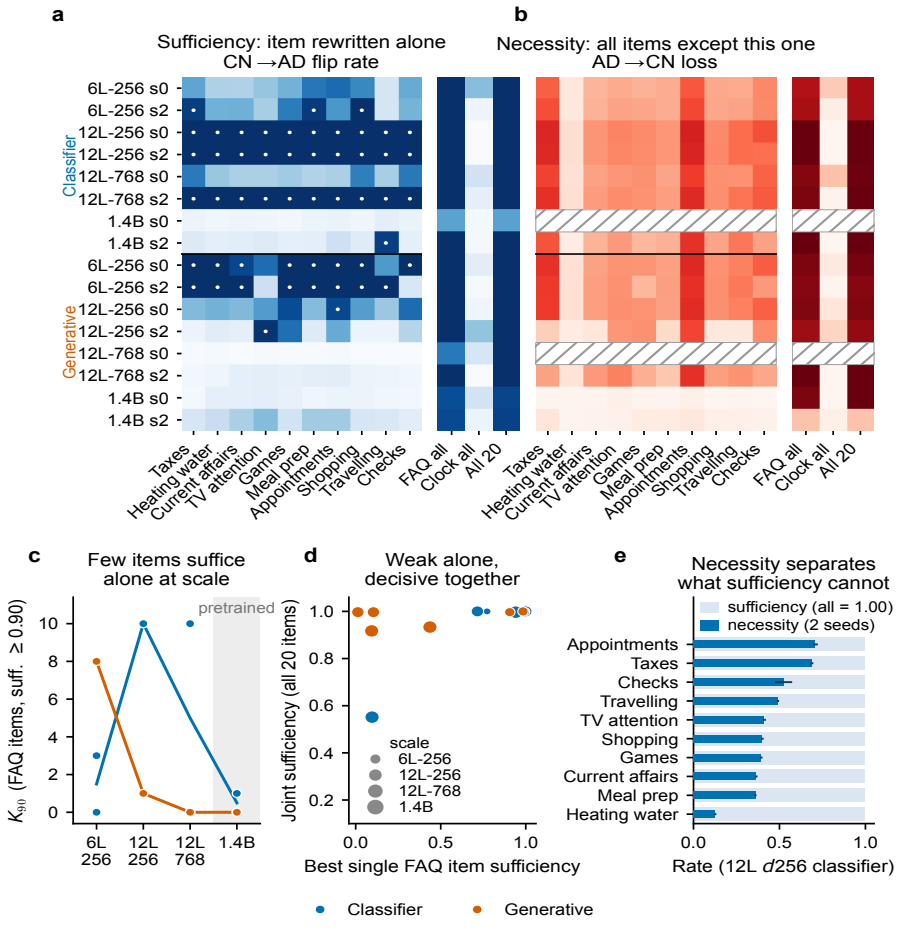
The impairment-inducing direction is less stable: one-item substitutions often preserve the flip rate, with only a small number of exceptions (Appendix D.6, Table 22). Thus, normalising interventions show recurring dependence on particular FAQ features, whereas impairment-inducing sets vary more across models and analysed runs. Because each intervention is evaluated on a model-specific cohort of correctly classified participants, these differences should be interpreted as within-model response patterns, not as paired causal effects of scale. They therefore reveal heterogeneity in how models aggregate functional evidence without establishing a shared representation of normal cognition or a common evidence-aggregation mechanism.

3.4 Interpretation and Scope of the Mechanistic Comparisons

The Pythia–scratch comparisons contrast complete training regimes rather than pretraining in isolation. Most mechanistic and counterfactual analyses are carried out on two training runs per configuration, and intervention effects are defined on model-specific eligible cohorts. We use "causal" strictly to refer to interventions on the model and its inputs, not to patient-level causal effects, and the positive cohort is broader than Alzheimer's disease alone. These constraints narrow the scope of inference, but they do not invalidate the within-task mechanistic comparisons. The matched models share the same task and principal architectural dimensions, and are evaluated using common data splits, matching procedures, intervention definitions, reliability criteria, and controls. The comparisons therefore remain valid for the question studied here: whether models with similar IID performance can nevertheless rely on different internal computations and different input evidence. They do not, however, isolate pretraining alone or establish that the observed mechanisms generalise

beyond the evaluated models and cohort. Full limitations and their analysis-specific implications are discussed in Appendix D.7.

Figure 4: **Individually weak, jointly decisive.** **a**, Sufficiency: CN→AD flip rate when one verbal item is rewritten (italic: block rewrites; dots: $\mathrm{Suff}_j \geq 0.90$, counted in K_{90}). **b**, Necessity $\mathrm{Nec}_j = F_{\mathrm{all}} - F_{-j}$ (AD→CN); hatched: ceiling too low to define. **c**, K_{90} falls with scale; the formulation ordering reverses with depth. **d**, Best single-item vs. joint (all-20) sufficiency; **e**, 12L/d256 classifier: all items fully sufficient, but necessity separates them. Single-item and joint rates use different conditioning.



4 Conclusion

Similar IID performance does not imply mechanistic interchangeability. In this ADNI-derived task, models with comparable macro-F1 scores differ in their sensitivity to numerical and verbal features, while architecture-matched training regimes differ in attention-head compressibility and in the depth at which attention-mediated influence peaks. FAQ descriptors consistently emerge as influential inputs, but their sufficiency, context-specific necessity, and sensitivity to wording vary across models and analysed runs. Score-preserving wording changes can also alter predictions, even in larger pretrained transformers such as Pythia-1.4B. These findings motivate evaluating models, cohorts, and interventions separately; establishing shared clinical reasoning mechanisms or general effects of scale will require further controlled comparisons.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to identify and review relevant literature, provide feedback on research methodology and experimental design, assist with method implementation and software code, support interpretation of experimental results and refinement of research hypotheses and claims, and improve manuscript text. Generative AI was also used to produce an initial version of a schematic figure, which was subsequently revised, post-processed, and finalised by the authors. All AI-assisted content and interpretations were independently reviewed against the underlying experimental results by the authors, and all methodological decisions and final scientific conclusions were made by the authors. Literature identified with AI assistance was verified against the original sources, and AI-generated or modified code was reviewed and tested for correctness. We did not use AI outputs as evidentiary sources or to generate the reported experimental data. The authors take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Task definitions, dataset construction, model architectures, training hyperparameters, checkpoint selection, and evaluation procedures are described in Appendix B. Detailed protocols for the mechanistic and causal-intervention analyses are provided in Appendices C and D, respectively. We report the training seeds, model-specific learning rates, intervention cohorts, reliability gates, and analysis criteria used for the reported results. An anonymised version of the code used to train, evaluate, and analyse the models is available at `https://anonymous.4open.science/r/Mech_Interpr_AD-EF0D/`.

REFERENCES

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), June 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Xuexin Chen, Ruichu Cai, Zhengting Huang, Yuxuan Zhu, Julien Horwood, Zhifeng Hao, Zijian Li, and José Miguel Hernández-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6486–6502. PMLR, 2024. URL `https://proceedings.mlr.press/v235/chen24d.html`.

Yingjie Feng, Jun Wang, Xianfeng Gu, Xiaoyin Xu, and Min Zhang. Large language models improve Alzheimer's disease diagnosis using multi-modality data. https://arxiv.org/abs/2305.19280v1, May 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Siva Rajesh Kasa, Karan Gupta, Sumegh Roychowdhury, Ashutosh Kumar, Yaswanth Biruduraju, Santhosh Kumar Kasa, Pattisapu Nikhil Priyatam, Arindam Bhattacharya, Shailendra Agarwal, and Vijay Huddar. Generative or discriminative? revisiting text classification in the era of transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9604–9626. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.486.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024.

Jian Lyu, Jingfeng Xue, Weizhi Meng, Weijie Han, Junbao Chen, and Zeyang Liu. Systematic design choices for fine-tuning text-classification models: Projection space, task instructions, and label encoding. *Applied Soft Computing*, 187:114341, 2026. doi: 10.1016/j.asoc.2025.114341.

Michail Mamalakis, Tiago Azevedo, Cristian Cosentino, Chiara D'Ercoli, Subati Abulikemu, Zhongtian Sun, Richard Bethlehem, and Pietro Lio. A monosemantic attribution framework for stable interpretability in clinical neuroscience large language models. *arXiv preprint arXiv:2601.17952*, 2026.

Gad A. Marshall, Amy S. Zoller, Natacha Lorius, Rebecca E. Amariglio, Joseph J. Locascio, Keith A. Johnson, Reisa A. Sperling, Dorene M. Rentz, and Alzheimer's Disease Neuroimaging Initiative. Functional activities questionnaire items that best discriminate and predict progression from clinically normal to mild cognitive impairment. *Current Alzheimer Research*, 12(5): 493–502, 2015. doi: 10.2174/1567205012050526115003. URL https://pubmed.ncbi.nlm.nih.gov/26017560/.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2407.08734.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nzpLWnVAyah.

R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74(3):201–209, January 2010. ISSN 1526-632X. doi: 10.1212/WNL.0b013e3181cb3e25.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. Llm circuit analyses are consistent across training and scale. *Advances in Neural Information Processing Systems*, 37:40699–40731, 2024.

Maria Eleftheria Vlontzou, Maria Athanasiou, Kalliopi V. Dalakleidi, Ioanna Skampardoni, Christos Davatzikos, and Konstantina Nikita. A comprehensive interpretable machine learning framework for mild cognitive impairment and alzheimer's disease diagnosis. *Scientific Reports*, 15:8410, 2025. doi: 10.1038/s41598-025-92577-6. URL https://www.nature.com/articles/s41598-025-92577-6.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63088–63112. PMLR, 2025.

Ying Wang, Lorelle Bowditch, Charlotte J. Molloy, Yinghua Yu, Peter Hibbert, and Farah Magrabi. Fine-tuning and evaluating large language models for patient safety tasks: Classification of contributing factors in incident reports. *Journal of the American Medical Informatics Association*, 33 (8):1494–1503, 2026. doi: 10.1093/jamia/ocag068.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12188–12200. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.679.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

A Extended Background and Related Work

Discriminative and generative formulations for classification. Transformer classification can be implemented by projecting a hidden state directly into a task-specific label space or by predicting a textual label through the language-model vocabulary. Recent systematic comparisons find that these choices can produce similar predictive performance while differing in sample efficiency, calibration, robustness, and training behaviour (Kasa et al., 2025; Lyu et al., 2026). Clinical studies have also compared encoder and decoder models primarily as predictive alternatives; for example, a recent patient-safety classification study found that a fine-tuned BERT encoder outperformed several decoder-only models on incident-report classification (Wang et al., 2026). Our comparison is orthogonal to the usual encoder-versus-decoder question: we hold the decoder-only backbone family fixed and vary the *task formulation* itself, asking whether a classification head and vocabulary-space label generation produce different causal computations.

Mechanisms across fine-tuning and scale. Mechanistic work increasingly asks whether circuits persist as models are trained, scaled, or adapted. Tigges et al. (2024) track circuits across Pythia, a decoder-only suite trained with the GPT-NeoX framework, and show substantial consistency across training and scale, while Wang et al. (2025) find that fine-tuning can preserve high-level circuit membership while substantially changing circuit connectivity. Fine-tuning can also alter access to capabilities learned during pretraining (Kotha et al., 2024). These studies motivate treating the full training history, namely pretraining together with downstream adaptation, as a mechanistic variable rather than only as a route to higher task accuracy.

Clinical interpretability and causal auditing. Clinical NLP has increasingly moved beyond performance toward explanations of model predictions. In Alzheimer's disease, Mamalakis et al. (2026) combine sparse-autoencoder representations with attribution methods to obtain more stable feature-level explanations. Attribution and decodability, however, do not establish that a feature is used by the model's own computation. We therefore combine representational analyses with interventions: circuit ablation, activation patching, feature-level activation swapping, and input-level counterfactuals. This distinction is especially important because LLMs can exploit shortcut features even when predictive accuracy is high (Yuan et al., 2024).

B Analytical Protocol

This appendix describes the analysis implemented in the accompanying codebase. We separate the experiments into two complementary groups. The first, *mechanistic approaches*, characterises what information is represented and how the diagnostic computation is organised. The second, *causal intervention and counterfactual analyses*, asks whether replacing, ablating, or rewriting a representation changes the model's own downstream decision. We use "causal" in this latter, within-model interventional sense; these experiments are not estimates of causal effects in the patient population.

B.1 Dataset

The neuropsychological measures from ADNI included: Clock Drawing and Copying Test (CDT), Auditory Verbal Learning Test (AVLT), Category Fluency for Animals (CFA), Trail Making Test, American National Adult Reading Test (ANART), and a Functional Assessment Questionnaire (FAQ). The cognitive numerical scores are embedded in this structured clinical summary.

The generated clinical text structure for each participant is:
Patient ID, Label, Generated Text
N, N, "The participant's sex is Male/Female. Their Date of Birth is MM/YYYY. Their Year of Birth is YYYY. Their Handedness is Right/Left. Their Marital status at baseline is Married/Widowed/Divorced/Never married/Unknown. Their education in years is N. Participant Retired? Yes/No. Type of Participant residence House/Condo/Apartment/Mobile Home/Retirement Community/Assisted Living/Skilled Nursing Facility/Other (owned or rented). Language to be used for testing the Participant English/Spanish. Participant's Primary Language English/Spanish/Other. The participant's Ethnicity is Hispanic or Latino/Not Hispanic or Latino/Unknown. Race American Indian or Alaskan Native/Asian/Native Hawaiian or Other Pacific Islander/Black or African Ameri-

can/White/More than one Race/Unknown. Information Source Participant Visit/Telephone Call. The participant's weight is N. The weight was measured in pounds/kilograms. The participant's Systolic - mmHg N. The participant's Diastolic - mmHg N. The participant's Seated Pulse Rate (per minute) is N. The participant's Respirations (per minute) are N. The participant's Temperature is N. The Temperature Source was Oral/Tympanic/Other. The Temperature Units were Fahrenheit/Celsius. Depressive symptoms present? Yes/No. On the Clock Drawing Test the participant answered the following questions in this way: Approximately circular face Correct/Incorrect. Symmetry of number placement Correct/Incorrect. Correctness of numbers Correct/Incorrect. Presence of the two hands Correct/Incorrect. Presence of the two hands, set to ten after eleven Correct/Incorrect. Clock Drawing Test: Total Score N. On the Clock copying task the participant scored as follows: Approximately circular face Correct/Incorrect. Symmetry of number placement Correct/Incorrect. Correctness of numbers Correct/Incorrect. Presence of the two hands Correct/Incorrect. Presence of the two hands, set to ten after eleven Correct/Incorrect. Clock copying task: Total Score N. On the Auditory Verbal Learning Test the participant scored as follows in each trial: Trial 1 Total N. Total Intrusions N. Trial 2 Total N. Total Intrusions N. Trial 3 Total N. Total Intrusions N. Trial 4 Total N. Total Intrusions N. Trial 5 Total N. Total Intrusions N. Trial 6 Total N. Total Intrusions N. List B Total N. Total Intrusions N. On the Auditory Verbal Learning Test the participant scored as follows: 30 Minute Delay Total N. Total Intrusions N. Recognition Score N. Total Intrusions N. On the Category Fluency Test Animals the scores were: - Total Correct N. Perseverations N. Intrusions N. Trail Making Test: Part A - Time to Complete N. Errors of Commission N. Errors of Omission N. Part B - Time to complete N. Errors of Commission N. Errors of Omission N. American National Adult Reading Test: ANART Total Score (Total # of errors) N. For the Functional Activities Questionnaire the participant scored as follows for each question: Writing checks, paying bills, or balancing checkbook. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Assembling tax records, business affairs, or other papers. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Shopping alone for clothes, household necessities, or groceries. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Playing a game of skill such as bridge or chess, working on a hobby. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Heating water, making a cup of coffee, turing off the stove. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Preparing a balanced meal. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Keeping track of current events. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Paying attention to and understanding a TV program, book, or magazine. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Remembering appointments, family occasions, holidays, medications. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Trail Making Test for FAQ score: Traveling out of the neighborhood, driving, or arranging to take public transportation. Normal (0)/Has difficulty, but does by self (1)/Requires assistance (2)/Dependent (3)/Never did, but could do now (0)/Never did, would have difficulty now (1). Total Score for FAQ is N. End of report."

B.2 Common data and read-out protocol

All analyses use as input the tokenized representation of the clinical narratives presented in the text subsection. The preprocessing and extraction of the structured clinical variables into text follow the general principles established in the ADNI clinical-language modelling pipeline of Mamalakis et al. (2026). A constant `` End of report.'' suffix is appended to every clinical narrative so that the classifier read-out does not coincide with the FAQ total-score token. The suffix is identical for every participant and therefore carries no label information. The maximum sequence length is 1,024 tokens.

The repository uses label 1 for the clinically impaired class and label 0 for cognitively normal (CN). The clinically impaired class includes Alzheimer's disease, Early Mild Cognitive Impairment, and

Late Mild Cognitive Impairment (Table 2). The control group includes Cognitively Normal and people with Subjective Memory Impairment. All measurements are taken at the baseline visit.

Table 2: Cohort composition. Mild cognitive impairment class (MCI) includes early and late presentations, of which some early MCI with diagnostic reversions are included in the control group. Labels 0 and 1 are balanced across train/val/test sets.

Split	Class	n	CN	MCI	AD	MCI share
Pretrained Pythia (LoRA): train / val / test						
train	CN (label 0)	513	508	5	0	1.0%
train	impaired (label 1)	514	0	392	122	76.3%
val	CN (label 0)	58	58	0	0	0.0%
val	impaired (label 1)	57	0	42	15	73.7%
test	CN (label 0)	571	567	4	0	0.7%
test	impaired (label 1)	571	0	419	152	73.4%
From-scratch GPT-NeoX: train / test						
train	CN (label 0)	571	566	5	0	0.9%
train	impaired (label 1)	571	0	434	137	76.0%
test	CN (label 0)	571	567	4	0	0.7%
test	impaired (label 1)	571	0	419	152	73.4%

The two task formulations have distinct task-specific read-outs. For the **classifier**, the bare clinical narrative is supplied to the decoder and a two-class score head reads the final non-padding position. For the **generative** formulation, the narrative is embedded in the training prompt and the model is trained to emit the textual class label. At evaluation time, we avoid stochastic decoding and compare the two answer-token logits at the final prompt position,

$$d(x) = z_1(x) - z_0(x). \tag{1}$$

where $z_1(x)$ is the logit of the token for the positive class (AD) and $z_0(x)$ is the logit of the token for the negative class (CN). Thus, all continuous generative-model analyses are defined on the same diagnostic margin used for classification.

B.3 Training, learning-rate selection, and saved-checkpoint evaluation

Pythia checkpoints are adapted with LoRA on all linear modules. The standard configuration uses rank $r = 32$, $\alpha = 64$, zero LoRA dropout, three epochs, AdamW with weight decay 0.01, a linear learning-rate schedule, five warm-up steps, per-device batch size 2, and gradient accumulation over 8 steps. The training seed changes optimisation and data order, whereas the data split is held fixed by dataset seed 42. There are no hyperparameters that require tuning for the from-scratch GPT-NeoX, as we used a fixed learning rate. While training is performed at bf16, all evaluations of the checkpoints and downstream analyses are performed at fp32.

Learning rate is tuned *per model scale and task formulation*. A fixed learning rate across the 14M–1.4B ladder is not assumed to be fair because the LoRA optimisation regime changes substantially with model scale. The primary selection procedure uses a fixed 10% validation split of the training data and selects the learning rate that maximises validation macro-F1 of the AD/CN *decision*. For the generative formulation, learning-model loss is not used as the selection criterion because it includes response-generation behaviour beyond the binary diagnostic decision.

The architecture-matched GPT-NeoX controls are trained from random initialisation for three epochs with batch size 2, gradient accumulation 8, warm-up 5 steps, weight decay 0.01, and learning rate 5×10^{-4}. Their principal shapes are 6L/$d = 256$/$h = 8$ and 12L/$d = 768$/$h = 12$, matching Pythia-31M and Pythia-160M respectively; the 12L/$d = 256$/$h = 8$ model is a depth-control configuration.

All headline metrics are recomputed after saving the checkpoint. LoRA adapters are merged into the base model, the saved checkpoint is reloaded in float32 and evaluation mode, and the full 1,142-example IID test set is scored. Classifier predictions are the argmax of the two-class score head; generative predictions are the sign of Eq. 1. We report accuracy, class-specific F1, macro-F1, sensitivity, specificity, and the predicted-positive fraction. A run is operationally labelled collapsed

Table 3: Training hyperparameters. Panel (a): every setting that fixes a trained model, for the two families the paper compares — LoRA fine-tunes of pretrained Pythia and architecture-matched GPT-NeoX models trained from scratch on the same corpus. Optimiser, schedule, batch size, epochs, sequence length and precision are identical across families, so the families differ only in initialisation, adaptation and learning rate. Panel (b): the learning rate per rung, the one setting that varies within the ladder; it was chosen per rung and per head on a held-out validation split, because a single rate applied across three orders of magnitude of scale leaves the small rungs at their initialisation. A rate in parentheses is a later, denser re-sweep, used only for the learning-rate robustness check, never for a shipped checkpoint, and it leaves every model carried into the interpretability analyses unchanged. *Seeds analysed*: the training seed moves initialisation, shuffling and dropout only; the data split is held at seed 42 and does not move with it, so repeated seeds measure run-to-run variance on one fixed split. Two of the five are carried through the interpretability pipeline in full, as independent replications.

Setting	Fine-tuned Pythia (LoRA)	From-scratch GPT-NeoX
Model and adaptation		
Initialisation	pretrained EleutherAI/pythia-{14M…1.4B}	random init (Pythia-160m tokenizer)
Sizes / shapes	6 rungs: 14M, 31M, 70M, 160M, 410M, 1.4B	6L/*d*256/8H, 12L/*d*256/8H, 12L/*d*768/12H
Adaptation	LoRA on all linear modules	full training, all parameters
LoRA rank r / α / dropout	32 / 64 / 0.0	—
Optimisation		
Learning rate	per rung, swept	5×10^{-4}
Scheduler	linear decay, 5 warmup steps	linear decay, 5 warmup steps
Optimiser	AdamW, weight decay 0.01	AdamW, weight decay 0.01
Batch size	2×8 accum. $= 16$	2×8 accum. $= 16$
Epochs (optimiser steps)	3 (192)	3 (213)
Max sequence length	1024 tokens	1024 tokens
Precision	bf16	bf16
Data and checkpoint selection		
Train / val / test	1027 / 115 / 1142	1142 / — / 1142
Checkpoint kept	best val macro-F_1	final epoch
Seed protocol		
Split seed	42, fixed	42, fixed
Training seeds run	0–4 (5 per rung per head)	0–4 (5 per shape per head)
Seeds analysed	0 and 2	0 and 2

when the predicted-positive fraction falls outside $[0.05, 0.95]$. Because some Pythia conditions are bimodal across seeds, per-seed results and collapse counts are accompanied by mean$\pm$s.d. summaries.

B.4 IID PREDICTIVE RESULTS

Learning-rate sensitivity and intermediate-scale degeneracy. We initially trained all model configurations using a fixed learning rate of 5×10^{-4}. Under this shared learning-rate setting (Table 4), the architecture-matched from-scratch controls were highly stable across seeds, with macro-F1 remaining near 0.82–0.83, whereas several intermediate-scale Pythia models exhibited substantial seed-to-seed variability and, in some cases, near-complete single-class collapse. The strongest example was the Pythia-160M generative model, for which all five runs were degenerate, yielding macro-F1 0.333 ± 0.000. Pythia-31M and Pythia-70M also showed large variance and class-imbalanced prediction behaviour, while performance recovered at the larger 410M and 1.4B scales.

Because a single learning rate need not provide comparable optimisation conditions across models that differ substantially in scale and task formulation, we performed a model-specific learning-rate sweep and retrained the Pythia ladder using the selected learning rate (Table 4) for each configuration.

Table 4: Learning-rate sweep. *Used* is the rate at which the checkpoints reported in this paper were trained. *Swept* lists the per-rung candidates. *Healthy* counts non-collapsed seeds at the selected rate.

Model	Formulation	Used	Swept	Healthy
Pretrained and LoRA-adapted Pythia				
14M	classifier	1×10^{-4}	–	5/5
14M	generative	1×10^{-4}	–	5/5
31M	classifier	3×10^{-4}	3×10^{-4}	5/5
31M	generative	3×10^{-4}	3×10^{-4}	4/5
70M	classifier	7×10^{-5}	$3 \times 10^{-5}, 5 \times 10^{-5}, 7 \times 10^{-5}, 2 \times 10^{-4}$	4/5
70M	generative	1×10^{-4}	$1 \times 10^{-4}, 1.5 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}$	5/5
160M	classifier	2×10^{-4}	2×10^{-4}	5/5
160M	generative	7×10^{-5}	$5 \times 10^{-5}, 7 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}$	3/5
410M	classifier	5×10^{-5}	–	5/5
410M	generative	5×10^{-5}	–	5/5
1.4B	classifier	1×10^{-4}	–	5/5
1.4B	generative	3×10^{-4}	–	5/5
Architecture-matched from-scratch GPT-NeoX controls				
6L, $d = 256$, $h = 8$	classifier	5×10^{-4}	–	5/5
6L, $d = 256$, $h = 8$	generative	5×10^{-4}	–	5/5
12L, $d = 256$, $h = 8$	classifier	5×10^{-4}	–	5/5
12L, $d = 256$, $h = 8$	generative	5×10^{-4}	–	5/5
12L, $d = 768$, $h = 12$	classifier	5×10^{-4}	–	5/5
12L, $d = 768$, $h = 12$	generative	5×10^{-4}	–	5/5

The resulting performance is reported in Table 6. Learning-rate tuning substantially reduced the degeneracy observed at the intermediate scales. In particular, Pythia-31M classifier improved from macro-F1 0.604 ± 0.180 to 0.803 ± 0.024, with all five seeds yielding non-collapsed predictions, while Pythia-31M generative improved from 0.551 ± 0.214 to 0.718 ± 0.216, with four of five seeds remaining non-degenerate. The Pythia-70M generative model similarly improved from 0.430 ± 0.189 to 0.680 ± 0.101, with all five seeds classified as healthy under the pre-specified collapse criterion.

The effect was more limited at 160M. The classifier remained non-degenerate across all seeds and showed reduced variability after tuning (0.697 ± 0.084 macro-F1), whereas the generative model improved from a fully degenerate 0.333 ± 0.000 to 0.468 ± 0.160, but only three of five seeds produced non-collapsed prediction distributions. Thus, the severe intermediate-scale failure observed under the shared learning rate is partly an optimisation effect rather than a fixed property of the architecture or task. Nevertheless, the residual instability of the 160M generative model indicates that adaptation remains substantially more fragile in this regime. The from-scratch controls remain stable across all configurations, providing an important reference point for the subsequent mechanistic comparisons.

Table 5: IID test performance across seeds, reported as mean ± s.d. Fine-tuned Pythia models use LoRA adaptation; from-scratch GPT-NeoX controls are trained directly on the diagnostic task. Generative models are evaluated from the AD/CN answer-token logit difference. † marks a rung in which at least one seed is a single-class collapse (predicted-AD outside $[0.05, 0.95]$). The mean and s.d. are over ALL seeds including those, so a dagger-marked mean lies between two modes rather than describing a spread around one, and its s.d. is not a standard error. Per-rung healthy counts and the learning rate each rung was trained at are in the full table. Macro-F1 is the primary metric because several intermediate-scale Pythia runs collapse toward a single predicted class.

Model	Formulation	Macro F1	Sens	Spec	Acc	$F1_{AD}$	pred-AD
Fine-tuned Pythia (LoRA)							
14M	classifier	0.82 ± 0.00	0.74 ± 0.02	0.90 ± 0.02	0.82 ± 0.00	0.80 ± 0.01	0.42 ± 0.02
14M	generative	0.81 ± 0.01	0.78 ± 0.05	0.84 ± 0.07	0.81 ± 0.01	0.80 ± 0.01	0.47 ± 0.06
31M	classifier	0.80 ± 0.02	0.73 ± 0.04	0.88 ± 0.03	0.80 ± 0.02	0.79 ± 0.03	0.42 ± 0.03
31M	generative†	0.72 ± 0.19	0.81 ± 0.10	0.70 ± 0.35	0.75 ± 0.13	0.78 ± 0.06	0.56 ± 0.22
70M	classifier†	0.63 ± 0.15	0.79 ± 0.20	0.55 ± 0.34	0.67 ± 0.11	0.70 ± 0.09	0.62 ± 0.26
70M	generative	0.68 ± 0.09	0.49 ± 0.21	0.93 ± 0.11	0.71 ± 0.06	0.60 ± 0.16	0.28 ± 0.15
160M	classifier	0.70 ± 0.08	0.68 ± 0.19	0.75 ± 0.25	0.72 ± 0.07	0.69 ± 0.08	0.46 ± 0.22
160M	generative†	0.47 ± 0.14	0.76 ± 0.24	0.35 ± 0.40	0.56 ± 0.09	0.62 ± 0.04	0.71 ± 0.32
410M	classifier	0.78 ± 0.07	0.67 ± 0.16	0.91 ± 0.05	0.79 ± 0.06	0.75 ± 0.12	0.38 ± 0.10
410M	generative	0.81 ± 0.01	0.75 ± 0.05	0.88 ± 0.04	0.81 ± 0.01	0.80 ± 0.02	0.43 ± 0.05
1.4B	classifier	0.83 ± 0.01	0.80 ± 0.04	0.86 ± 0.04	0.83 ± 0.01	0.82 ± 0.01	0.47 ± 0.04
1.4B	generative	0.80 ± 0.02	0.79 ± 0.06	0.81 ± 0.05	0.80 ± 0.02	0.80 ± 0.02	0.49 ± 0.05
From-scratch GPT-NeoX controls							
6L/$d = 256/h = 8$	classifier	0.83 ± 0.00	0.79 ± 0.02	0.86 ± 0.03	0.83 ± 0.00	0.82 ± 0.00	0.46 ± 0.02
6L/$d = 256/h = 8$	generative	0.83 ± 0.00	0.77 ± 0.02	0.88 ± 0.02	0.83 ± 0.00	0.82 ± 0.01	0.45 ± 0.02
12L/$d = 256/h = 8$	classifier	0.83 ± 0.00	0.77 ± 0.00	0.90 ± 0.00	0.83 ± 0.00	0.82 ± 0.00	0.44 ± 0.00
12L/$d = 256/h = 8$	generative	0.83 ± 0.01	0.77 ± 0.01	0.88 ± 0.01	0.83 ± 0.01	0.82 ± 0.01	0.45 ± 0.01
12L/$d = 768/h = 12$	classifier	0.82 ± 0.01	0.78 ± 0.02	0.86 ± 0.03	0.82 ± 0.01	0.82 ± 0.01	0.46 ± 0.02
12L/$d = 768/h = 12$	generative	0.82 ± 0.01	0.77 ± 0.01	0.88 ± 0.02	0.83 ± 0.01	0.82 ± 0.01	0.45 ± 0.01

Table 6: Effect of validation-based learning-rate selection, for the six rungs the sweep moved. The other six (14M and 410M both heads, 1.4B both heads) kept their original learning rate and are omitted. Rates were chosen per rung and per head on a held-out validation split, never on the test set reported here. *collapsed* counts seeds that predict a single class for every patient; such a seed scores macro-F1 0.333 by construction, so a rung's mean lies between its working and collapsed seeds. Selection rescues 9 of the 13 collapsed seeds, which is most of what the gain consists of – the largest improvement, 70M generative at $+0.250$, comes with 4 fewer collapsed seeds. All values are macro-F1 on the IID test split, mean ± s.d. over five training seeds.

		learning-rate untuned			learning-rate tuned			
Model	Variant	LR	macro-F1	collapsed	LR	macro-F1	collapsed	Δ
31M	clf	1×10^{-4}	0.604 ± 0.180	1/5	3×10^{-4}	0.803 ± 0.024	0/5	$+0.199$
	gen	1×10^{-4}	0.551 ± 0.214	2/5	3×10^{-4}	0.718 ± 0.216	1/5	$+0.167$
70M	clf	1×10^{-4}	0.626 ± 0.184	1/5	7×10^{-5}	0.631 ± 0.173	1/5	$+0.006$
	gen	3×10^{-4}	0.430 ± 0.189	4/5	1×10^{-4}	0.680 ± 0.101	0/5	$+0.250$
160M	clf	3×10^{-4}	0.684 ± 0.189	0/5	2×10^{-4}	0.697 ± 0.094	0/5	$+0.014$
	gen	3×10^{-4}	0.333 ± 0.000	5/5	7×10^{-5}	0.468 ± 0.160	2/5	$+0.135$

C Mechanistic Approaches

The analyses in this section characterise the organisation and availability of diagnostic information. Circuit extraction includes an intervention, but its primary role here is to quantify the compressibility of the attention computation. Direct attribution, attention visualisation, and linear probing are descriptive and are not, by themselves, treated as evidence that a feature or component is causally used. However, they provide guidance for the targeted intervention experiments presented in the next section.

C.1 GREEDY ATTENTION-HEAD CIRCUIT EXTRACTION

We extract an attention-head subset using greedy mean ablation. For example i with label $y_i \in \{0, 1\}$, the signed correct-class margin is

$$D_i(M) = (2y_i - 1)\,[z_1(M, x_i) - z_0(M, x_i)] = z_{\text{correct}} - z_{\text{incorrect}}. \tag{2}$$

Signing by the true label prevents cancellation between classes in a balanced cohort.

For attention head (ℓ, h), mean ablation replaces the head output by its mean activation over the circuit-fit examples and token positions. Let H be the intact set of heads, C a candidate circuit, and $\varnothing$ the "empty" attention circuit in which all attention heads are mean-ablated while embeddings, MLPs, and the task read-out remain intact. For a sample set S, define

$$\bar{D}(C; S) = \frac{1}{|S|} \sum_{i \in S} D_i(M(C)). \tag{3}$$

Circuit extraction is attempted only when the intact margin is measurable,

$$|t_{\text{base}}| = \left| \frac{\bar{D}(H; F)}{s_H / \sqrt{|F|}} \right| \geq 2, \tag{4}$$

and when the paired head-attributable range $D_i(H) - D_i(\varnothing)$ also satisfies $|t_R| \geq 2$. If either gate fails, the circuit size is reported as *undefined*.

Heads are ranked on the circuit-fit subset F by the absolute mean change caused by individually mean-ablating each head,

$$I_{\ell h} = \left| \frac{1}{|F|} \sum_{i \in F} [D_i(M(H \setminus \{(\ell, h)\})) - D_i(M(H))] \right|. \tag{5}$$

Heads are restored greedily in decreasing $I_{\ell h}$. Circuit faithfulness uses a two-point normalisation,

$$\Phi(C; S) = \frac{\bar{D}(C; S) - \bar{D}(\varnothing; S)}{\bar{D}(H; S) - \bar{D}(\varnothing; S)}. \tag{6}$$

The reported circuit is the smallest C_k satisfying

$$\Phi(C_k; F) \geq 0.90 \quad \text{and} \quad \Phi(C_k; V) \geq 0.90, \tag{7}$$

where V is not used to determine the head ranking. The current study uses $|F| = 50$ and $|V| = 200$ from a balanced 250-example draw. Both are sampled from the official IID test partition, so we call them *circuit-fit* and *circuit-held-out* subsets rather than dataset training and test sets. A more conservative final protocol can fit the ranking on the training/validation partition and reserve the official test split for circuit evaluation.

All final circuit analyses use float32 inference, signed margins, mean ablation, and the held-out faithfulness gate.

C.1.1 CIRCUIT-EXTRACTION RESULTS

The current circuit results show that attention-head concentration is not fixed by macro-architecture alone, and it is also seed-dependent.

On the same scale, we observe that for the 1.4B models, where the baseline margin is reliable, the circuit size of each head variant is seed-dependent: in seed 0, the classifier model is distributed (93.2%) and the generative model is concentrated (53.1%); whereas in seed 2, the classifier variant is concentrated (38.3%) and the generative model is more diffuse at (62.8%).

Overall, when considering macro-architecture scaling within each seed, the more distributed circuits appear at larger model sizes: the most distributed circuits appear for 1.4B and 160M models in both seeds 0 (160M 93.8%, 1.4B 93.2%) and 2 (160M 95.8%), whereas the most concentrated circuits belong to the from-scratch GPT-NeoX models, especially 12L d=768, h=12 (16% and 11%)

Even at the same size, training history affects the circuit size. Most strinkigly, in seed 2, 12L d=768 circuit sizes range from 4.2 to 8.3%, whereas the architecture-matched pretrained 160M models range from 56.9-95.8%, up to ten times bigger.

To ensure this is not a seed-dependent effect, we measure the average circuit size across 5 seeds in Table 9. Training history does not obviously affect circuit size in smaller 31M/6L models, but instead it drastically increases circuit size in the 12L/160M models. From scratch GPT-Neox models have more concentrated circuit sizes, and pretraining of larger models increases its circuit size, suggesting that it may be utilising its prevoius knowledge to solve the task at hand.

Thus, matching depth, hidden width, and head count does not determine circuit compressibility.

The current 160M generative Pythia for seed 0 do not support circuit-size estimation under the pre-specified reliability gate. Their intact signed margins is below the extraction threshold of 2. We therefore report these circuits as *undefined*. At 1.4B,

Table 7: Attention-head circuit concentration in architecture-matched from-scratch controls and the current fine-tuned Pythia checkpoints using Seed 0. *Circuit* reports attention heads retained of the total. For Pythia, $|t|$ is the reliability of the intact signed logit margin and *Faithfulness* is the held-out two-point circuit faithfulness. †When $|t| < 2$, or when the heads account for no measurable part of the margin, extraction is not performed and circuit size is undefined. ‡Rung retrained at the learning rate selected by the corrected sweep. The Pythia values should be regenerated if the final validation-selected learning rates change.

Model	Formulation	$\lvert t\rvert$	Circuit	% heads	Faithfulness
Architecture-matched from-scratch GPT-NeoX controls					
6L, $d = 256$, $h = 8$	classifier	7.05	26/48	54.2%	91.4%
6L, $d = 256$, $h = 8$	generative	6.73	17/48	35.4%	90.6%
12L, $d = 256$, $h = 8$	classifier	6.69	77/96	80.2%	90.7%
12L, $d = 256$, $h = 8$	generative	7.49	45/96	46.9%	90.1%
12L, $d = 768$, $h = 12$	classifier	6.32	23/144	16.0%	90.3%
12L, $d = 768$, $h = 12$	generative	6.97	16/144	11.1%	93.2%
Pretrained and LoRA-adapted Pythia					
14M	classifier	6.88	11/24	45.8%	91.4%
14M	generative	6.39	14/24	58.3%	95.2%
31M‡	classifier	6.83	36/48	75.0%	94.6%
31M‡	generative	7.00	32/48	66.7%	91.0%
160M‡	classifier	5.56	135/144	93.8%	92.5%
160M‡	generative	–	undefined†	–	–
1.4B	classifier	6.51	358/384	97.6%	93.2%
1.4B	generative	6.44	204/384	53.1%	90.2%

Table 8: Attention-head circuit concentration in architecture-matched from-scratch controls and the current fine-tuned Pythia checkpoints using Seed 2. *Circuit* reports attention heads retained of the total. For Pythia, $|t|$ is the reliability of the intact signed logit margin and *Faithfulness* is the held-out two-point circuit faithfulness. †When $|t| < 2$, or when the heads account for no measurable part of the margin, extraction is not performed and circuit size is undefined. ‡Rung retrained at the learning rate selected by the corrected sweep. The Pythia values should be regenerated if the final validation-selected learning rates change.

Model	Formulation	$\|t\|$	Circuit	% heads	Faithfulness
Architecture-matched from-scratch GPT-NeoX controls					
6L, $d = 256$, $h = 8$	classifier	7.43	13/48	27.1%	91.4%
6L, $d = 256$, $h = 8$	generative	6.51	10/48	20.8%	91.8%
12L, $d = 256$, $h = 8$	classifier	6.53	5/96	5.2%	90.3%
12L, $d = 256$, $h = 8$	generative	7.60	45/96	46.9%	90.0%
12L, $d = 768$, $h = 12$	classifier	6.23	6/144	4.2%	93.0%
12L, $d = 768$, $h = 12$	generative	7.05	12/144	8.3%	90.0%
Pretrained and LoRA-adapted Pythia					
14M	classifier	7.58	12/24	50.0%	94.6%
14M	generative	7.81	16/24	66.7%	93.0%
31M‡	classifier	7.97	26/48	54.2%	91.5%
31M‡	generative	6.22	17/48	35.4%	91.3%
160M‡	classifier	4.41	82/144	56.9%	90.2%
160M‡	generative	3.93	138/144	95.8%	93.4%
1.4B	classifier	6.83	147/384	38.3%	94.7%
1.4B	generative	6.35	241/384	62.8%	90.1%

Table 9: Architecture-matched circuit comparison across 5 seeds. *Seeds* is the number of extractions attempted and *Valid* the number passing both reliability gates: $|t| \geq 2$ on the intact signed logit margin, and $|t_R| \geq 2$ on the part of that margin the heads jointly carry. A refusal means circuit size is undefined. *Heads for 90%* and *Fraction* are mean $\pm$ s.d. over valid seeds only. **Runs excluded from the summaries:** Pythia-31M generative seed 4 (baseline gate, $|t| = 1.70$); Pythia-160M classifier seed 4 (head-range gate, $|t_R| = 1.96$); Pythia-160M generative seed 0 (head-range gate, $|t_R| = 1.67$); Pythia-160M generative seed 1 (baseline gate, $|t| = 0.25$); Pythia-160M generative seed 3 (baseline gate, $|t| = 0.44$); Pythia-160M generative seed 4 (head-range gate, $|t_R| = 1.54$).

Condition	Formulation	Seeds	Valid	Heads for 90%	Fraction
From-scratch 6L/$d256$	classifier	5	5	24.6 ± 7.5	0.51 ± 0.16
Pythia-31M	classifier	5	5	31.4 ± 10.2	0.65 ± 0.21
From-scratch 6L/$d256$	generative	5	5	18.4 ± 7.4	0.38 ± 0.15
Pythia-31M	generative	5	4	24.2 ± 11.2	0.51 ± 0.23
From-scratch 12L/$d768$	classifier	5	5	9.6 ± 7.5	0.07 ± 0.05
Pythia-160M	classifier	5	4	100.2 ± 30.0	0.70 ± 0.21
From-scratch 12L/$d768$	generative	5	5	26.6 ± 8.1	0.18 ± 0.06
Pythia-160M	generative	5	1	138.0	0.96

C.2 Direct logit attribution

Direct logit attribution asks which component outputs at the read-out position are aligned with the final class direction. For the generative model, the class direction is the difference between the two answer-token unembedding vectors; for the classifier it is the difference between the two columns of the trained score head. If $o_c(x)$ is the output of component c at the read-out position and $w_1 - w_0$ is the class direction,

$$A_c(x) = o_c(x)^\top (w_1 - w_0). \tag{8}$$

We compute this quantity for each attention head, each MLP block, and the input embedding, retain per-example values, and report averages across the analysed cohort. Because LayerNorm is nonlinear, the sum of direct attributions is used as a diagnostic rather than required to exactly equal the

observed logit margin. Raw attribution magnitude is not compared directly across model sizes; for cross-model comparisons we report within-model distributions, ranks, or normalised values.

C.2.1 DIRECT-ATTRIBUTION RESULTS

Direct-logit-attribution results are available for both the pretrained and from-scratch models, and they place their strongest heads in different parts of the network (Table 10, and corresponding Figures in Appendix C.2.2). Across the from-scratch controls, the top attention-head contributions sit almost exclusively in layers 0–1: the 6L/d=256 and 12L/d=768 models place their strongest heads at layer 0 or 1 in every seed and formulation, with concentrated shares (top heads carrying 15–41% in the wide model). The one departure is the deep-narrow 12L/d=256 model, where a single run (generative, seed 2) places its top heads at layers 3–5 while the remaining three runs stay at layers 0–1 — suggesting this configuration admits more than one circuit placement rather than a systematic shift with depth. The pretrained models, by contrast, concentrate mid-to-late at every scale: top heads span layers 2–5 (of 6) at 31M, 3–10 (of 12) at 160M, and 11–22 (of 24) at 1.4B, a consistently mid-to-late placement in relative depth. Layer predominance is otherwise stable across seeds, although the specific heads within those layers vary between seeds and between the AD and CN directions.

MLP directionality is heterogeneous across the set, and neither objective nor architecture predicts which member of a pair is concordant. The clearest single case is the wide generative control at seed 0, where all twelve MLPs write toward AD with nothing cancelled ($C = +1.00$, net $+41.99$), against a heavily cancelling classifier counterpart (12.23 vs 8.12, $C = +0.20$); the same polarity holds at 1.4B, whose classifier at seed 2 is the most cancelled run in the set ($C = +0.01$). But the assignment reverses at 12L/d=256 and 160M, where the classifier is the more concordant arm, and the 6L classifier is fully concordant in the opposite direction ($C = -1.00$, 0/6 MLPs AD-ward, at both seeds) while its generative counterpart cancels almost completely. Concordance is also seed-sensitive within rungs (the retrained 31M classifier spans $+0.13$ to $+0.83$; the wide generative model $+1.00$ to $+0.45$), so it is best read as a property of the individual solution rather than of the objective or the width. Two regularities do persist: the only fully concordant solutions occur in the from-scratch controls, with no LoRA-adapted rung exceeding $|C| = 0.83$, consistent with adaptation reorienting rather than replacing pretrained MLP function; and the 31M and 160M generative rungs net toward CN in every run. In the heavily cancelling regimes especially, these sums should be read as maps of direct output alignment rather than as evidence of causal sufficiency.

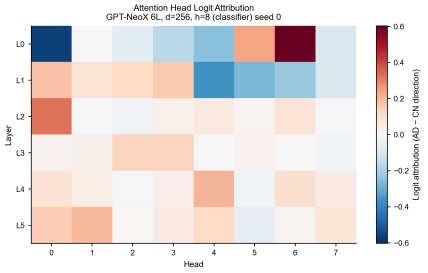
Table 10: Direct contribution of attention heads to the AD−CN logit difference at the read-out token (head output projected onto $W_U[:, \mathrm{AD}] - W_U[:, \mathrm{CN}]$). For each direction we list the 3 strongest heads, extending to 5 while the next head carries at least half the strongest; brackets give each head's share of that direction's total. Shares are used rather than raw logits because per-head scale varies with the model (0.07 logits at 31M vs. 16 in the 12L/d=768 controls). Heads are named LℓHh, strongest first. Unlike circuit extraction, attribution applies no reliability gate, so rows appear even where the intact margin is indistinguishable from zero. These heads form the shortlist for attention-pattern visualisation.

Model	Formulation	Seed	Dir.	Heads
Architecture-matched from-scratch GPT-NeoX controls				
6L, $d = 256$, $h = 8$	clf	0	AD	L0H6 [18%], L2H0 [10%], L0H5 [7%]
			CN	L0H0 [25%], L1H4 [16%], L1H5 [12%]
6L, $d = 256$, $h = 8$	clf	2	AD	L1H7 [9%], L0H3 [9%], L1H3 [8%], L1H1 [7%], L2H0 [6%]
			CN	L0H0 [25%], L1H4 [25%], L1H2 [24%], L1H0 [15%]
6L, $d = 256$, $h = 8$	gen	0	AD	L0H7 [24%], L0H0 [22%], L1H4 [13%]
			CN	L0H3 [29%], L0H4 [25%], L1H5 [11%]
6L, $d = 256$, $h = 8$	gen	2	AD	L0H2 [38%], L0H1 [15%], L1H1 [12%]
			CN	L0H4 [28%], L1H2 [25%], L0H5 [16%]
12L, $d = 256$, $h = 8$	clf	0	AD	L1H3 [12%], L1H2 [6%], L1H5 [5%]
			CN	L0H3 [11%], L3H3 [9%], L3H6 [9%], L2H1 [6%]
12L, $d = 256$, $h = 8$	clf	2	AD	L0H3 [6%], L1H7 [5%], L0H7 [3%]
			CN	L0H5 [29%], L1H1 [27%], L0H1 [18%]
12L, $d = 256$, $h = 8$	gen	0	AD	L0H7 [14%], L0H6 [10%], L0H3 [10%]
			CN	L1H4 [15%], L3H7 [10%], L5H4 [9%], L3H5 [8%]
12L, $d = 256$, $h = 8$	gen	2	AD	L3H5 [15%], L3H7 [9%], L3H3 [8%]
			CN	L4H3 [26%], L4H7 [8%], L5H1 [7%]
12L, $d = 768$, $h = 12$	clf	0	AD	L0H7 [24%], L0H8 [18%], L0H3 [15%], L1H7 [13%]
			CN	L0H11 [16%], L0H9 [15%], L0H5 [9%]
12L, $d = 768$, $h = 12$	clf	2	AD	L0H0 [22%], L0H3 [20%], L0H7 [19%]
			CN	L0H5 [41%], L0H8 [22%], L0H10 [13%]
12L, $d = 768$, $h = 12$	gen	0	AD	L0H8 [31%], L6H10 [9%], L8H4 [5%]
			CN	L0H5 [15%], L0H9 [13%], L1H7 [12%], L0H3 [11%], L10H0 [9%]
12L, $d = 768$, $h = 12$	gen	2	AD	L0H7 [27%], L1H5 [12%], L0H4 [11%]
			CN	L0H10 [31%], L0H1 [18%], L0H3 [14%]
Pretrained and LoRA-adapted Pythia				
31M	clf	0	AD	L2H6 [18%], L2H1 [12%], L0H1 [7%]
			CN	L3H5 [15%], L4H3 [15%], L1H3 [13%], L3H6 [8%], L4H5 [8%]
31M	clf	2	AD	L5H2 [24%], L3H5 [13%], L2H2 [11%]
			CN	L4H6 [20%], L2H3 [12%], L3H0 [10%]
31M	gen	0	AD	L4H5 [18%], L4H6 [18%], L3H0 [11%], L4H4 [10%]
			CN	L5H3 [21%], L5H6 [15%], L5H5 [10%]
31M	gen	2	AD	L4H1 [12%], L2H3 [10%], L3H5 [10%], L0H7 [9%], L4H5 [8%]
			CN	L5H6 [16%], L5H3 [14%], L5H1 [14%], L5H0 [8%]
160M	clf	0	AD	L7H11 [14%], L9H3 [6%], L9H1 [5%]
			CN	L5H7 [9%], L8H0 [7%], L6H7 [6%], L11H11 [5%], L6H4 [5%]
160M	clf	2	AD	L9H10 [13%], L11H7 [5%], L3H3 [5%]
			CN	L6H3 [6%], L11H10 [4%], L11H4 [4%], L11H0 [4%], L5H0 [3%]
160M	gen	0	AD	L3H2 [5%], L5H4 [5%], L5H5 [5%], L6H10 [4%], L3H0 [4%]
			CN	L10H1 [14%], L7H8 [7%], L11H7 [6%]
160M	gen	2	AD	L7H11 [7%], L9H8 [6%], L7H0 [4%], L8H1 [4%], L5H5 [4%]
			CN	L8H7 [12%], L9H4 [11%], L11H7 [7%], L10H7 [6%], L10H1 [6%]
1.4B	clf	0	AD	L22H0 [6%], L9H10 [2%], L22H2 [2%]
			CN	L11H5 [5%], L7H15 [3%], L22H14 [3%], L10H15 [3%]
1.4B	clf	2	AD	L11H1 [3%], L13H10 [3%], L23H6 [2%], L11H2 [2%], L12H5 [2%]
			CN	L21H4 [3%], L12H0 [3%], L12H7 [2%], L15H14 [2%], L22H13 [2%]
1.4B	gen	0	AD	L17H7 [9%], L17H13 [6%], L23H7 [3%]
			CN	L14H12 [12%], L10H0 [9%], L14H6 [4%]

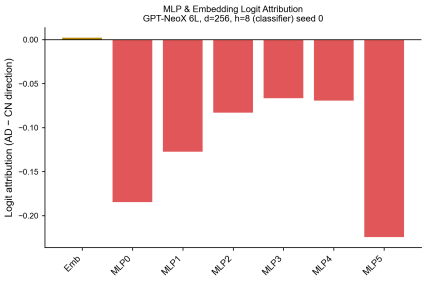
Table 11: MLP directionality in the logit attribution: whether a model's MLPs push the same way as one another, and which way. Each block's MLP output at the read-out token is projected onto $W_U[:, \mathrm{AD}] - W_U[:, \mathrm{CN}]$, positive toward AD and negative toward CN. $\Sigma_{\to \mathrm{AD}}$ and $\Sigma_{\to \mathrm{CN}}$ are the two opposing masses, each summed over the layers pushing that way, and Net is their difference. Concordance is $C = \mathrm{Net}/\mathrm{gross}$ with $\mathrm{gross} = \Sigma_{\to \mathrm{AD}} + \Sigma_{\to \mathrm{CN}}$: $C = \pm 1$ when every MLP pushes the same way and nothing is cancelled, and $C \approx 0$ when the two sides are of equal size and annihilate, the sign saying which way the survivor points. Cancellation, the share of the gross mass annihilated, is $1 - |C|$ and is therefore not given its own column. A ratio is reported rather than a sum of logits because the scale is a property of the model and not of the task: the gross MLP mass spans 0.29 to 42.0 logits across this set. The last column counts the layers contributing toward AD. Attention heads are the subject of the companion attribution table; embedding contributions are excluded. The pattern is not uniform across objectives. In the wide 12L/d=768 pair the generative model is the concordant one – all twelve of its MLPs point toward AD at seed 0 – and the classifier the cancelling one, as at 1.4B; but at 12L/d=256 and at 160M the classifier is the more concordant arm of the pair, and the 6L/d=256 classifier is fully concordant toward CN. Attribution applies no reliability gate, unlike circuit extraction, so a row appears even where the intact margin is not distinguishable from zero.

Model	Formulation	Seed	$\Sigma_{\to \mathrm{AD}}$	$\Sigma_{\to \mathrm{CN}}$	Net	C	MLPs AD-ward
Architecture-matched from-scratch GPT-NeoX controls							
6L, $d = 256$, $h = 8$	clf	0	0.00	0.76	-0.76	-1.00	0/6
6L, $d = 256$, $h = 8$	clf	2	0.00	0.64	-0.64	-1.00	0/6
6L, $d = 256$, $h = 8$	gen	0	0.33	0.37	-0.04	-0.06	2/6
6L, $d = 256$, $h = 8$	gen	2	0.92	0.83	$+0.09$	$+0.05$	2/6
12L, $d = 256$, $h = 8$	clf	0	2.13	0.46	$+1.66$	$+0.64$	10/12
12L, $d = 256$, $h = 8$	clf	2	2.06	0.24	$+1.82$	$+0.79$	8/12
12L, $d = 256$, $h = 8$	gen	0	0.57	0.70	-0.14	-0.11	7/12
12L, $d = 256$, $h = 8$	gen	2	0.69	2.69	-2.01	-0.59	4/12
12L, $d = 768$, $h = 12$	clf	0	12.23	8.12	$+4.11$	$+0.20$	9/12
12L, $d = 768$, $h = 12$	clf	2	7.86	4.26	$+3.59$	$+0.30$	8/12
12L, $d = 768$, $h = 12$	gen	0	41.99	0.00	$+41.99$	$+1.00$	12/12
12L, $d = 768$, $h = 12$	gen	2	5.99	2.26	$+3.72$	$+0.45$	7/12
Pretrained and LoRA-adapted Pythia							
31M	clf	0	0.22	0.17	$+0.05$	$+0.13$	3/6
31M	clf	2	0.26	0.02	$+0.24$	$+0.83$	5/6
31M	gen	0	0.19	0.45	-0.26	-0.41	2/6
31M	gen	2	0.10	0.35	-0.25	-0.54	3/6
160M	clf	0	4.52	5.21	-0.69	-0.07	7/12
160M	clf	2	1.84	1.35	$+0.49$	$+0.15$	7/12
160M	gen	0	3.12	3.27	-0.14	-0.02	5/12
160M	gen	2	1.49	2.30	-0.81	-0.21	4/12
1.4B	clf	0	6.45	3.91	$+2.54$	$+0.24$	19/24
1.4B	clf	2	7.73	7.55	$+0.19$	$+0.01$	10/24
1.4B	gen	0	15.78	4.53	$+11.25$	$+0.55$	18/24
1.4B	gen	2	11.22	3.90	$+7.32$	$+0.48$	13/24

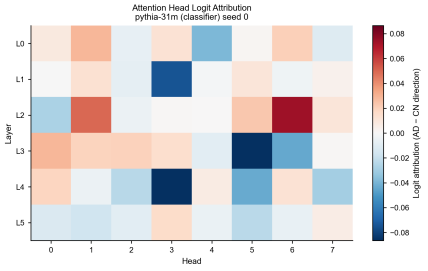
C.2.2 DIRECT-ATTRIBUTION FIGURES



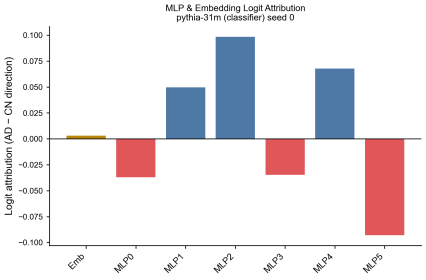
(a) Classifier 6-layer model Attention head logit attribution



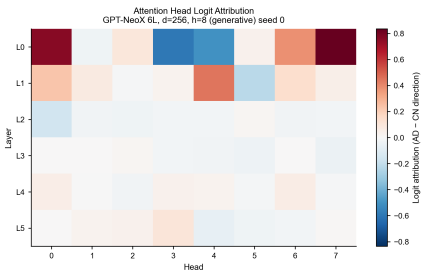
(b) Classifier 6-layer GPT-NeoX model MLP logit attribution



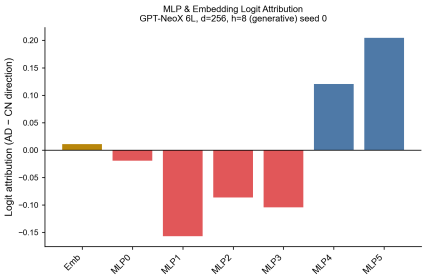
(c) Classifier 31M pretrained Pythia model Attention head logit attribution



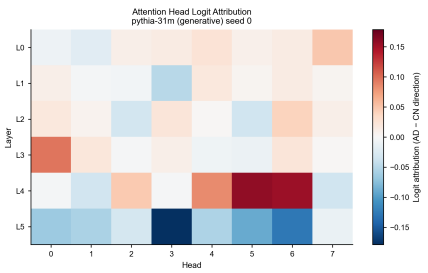
(d) Classifier 31M pretrained Pythia model Attention head logit attribution



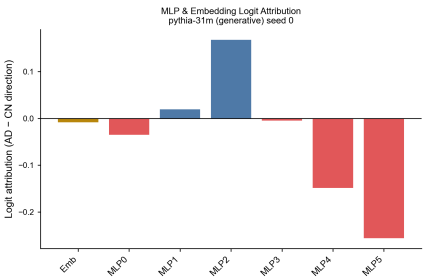
(e) Generative 6-layer model Attention head logit attribution



(f) Generative 6-layer GPT-NeoX model MLP logit attribution

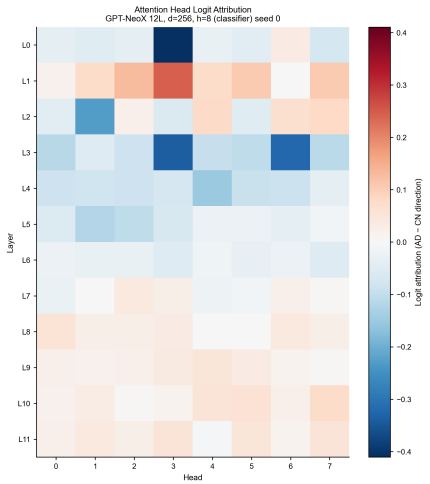


(g) Generative 31M pretrained Pythia model Attention head logit attribution

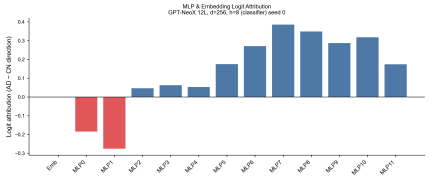


(h) Generative 31M pretrained Pythia model Attention head logit attribution

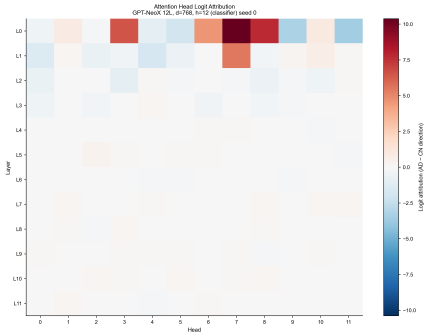
Figure 5: Logit attribution scores for Classifier and Generative 31M pretrained Pythia and 6L from-scratch GPT-NeoX Models at Seed 0.



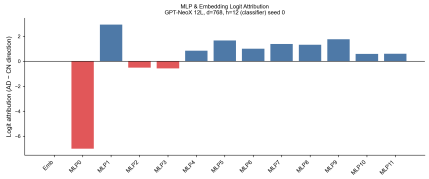
(a) Classifier 12-layer d=256 GPT-NeoX model Attention head logit attribution



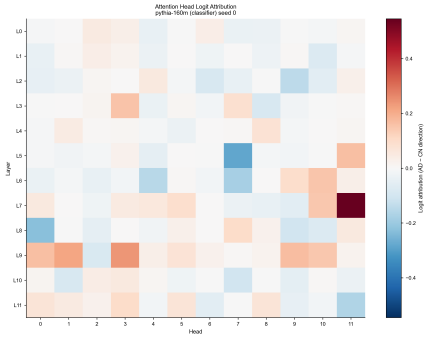
(b) Classifier 12-layer d=256 GPT-NeoX model MLP logit attribution



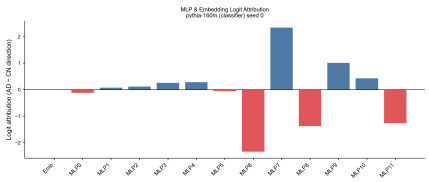
(c) Classifier 12-layer d=768 GPT-NeoX model Attention head logit attribution



(d) Classifier 12-layer d=768 GPT-NeoX model Attention head logit attribution

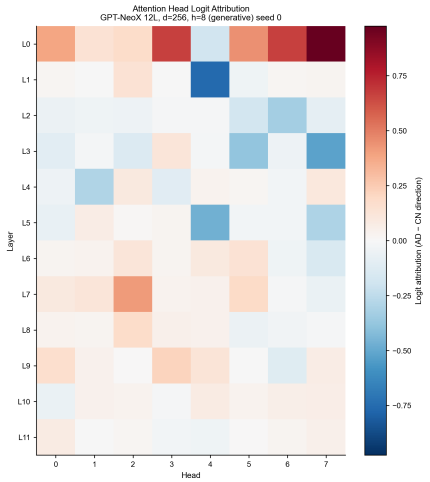


(e) Classifier 160M pretrained Pythia model Attention head logit attribution

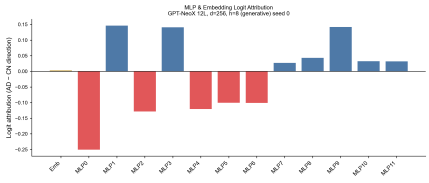


(f) Classifier 160M pretrained Pythia model Attention head logit attribution

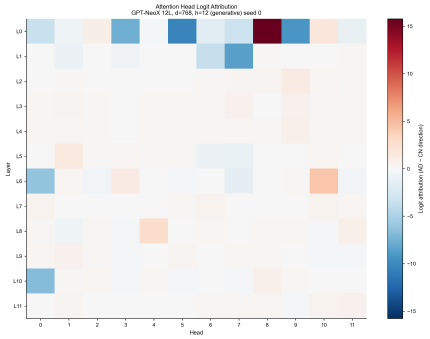
Figure 6: Logit attribution scores for Classifier 160M pretrained Pythia and 12L d=256 and d=768 from scratch GPT-NeoX models at Seed 0.



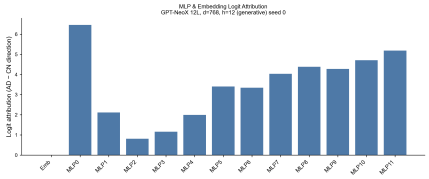
(a) Generative 12-layer d=256 GPT-NeoX model Attention head logit attribution



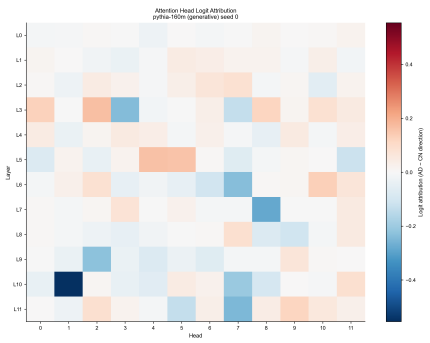
(b) Generative 12-layer d=256 GPT-NeoX model MLP logit attribution



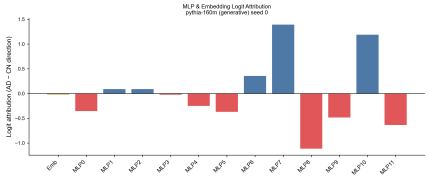
(c) Generative 12-layer d=768 GPT-NeoX model Attention head logit attribution



(d) Generative 12-layer d=768 GPT-NeoX model Attention head logit attribution

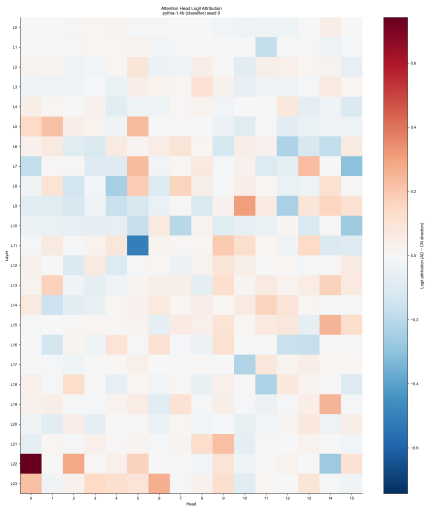


(e) Generative 160M pretrained Pythia model Attention head logit attribution



(f) Generative 160M pretrained Pythia model Attention head logit attribution

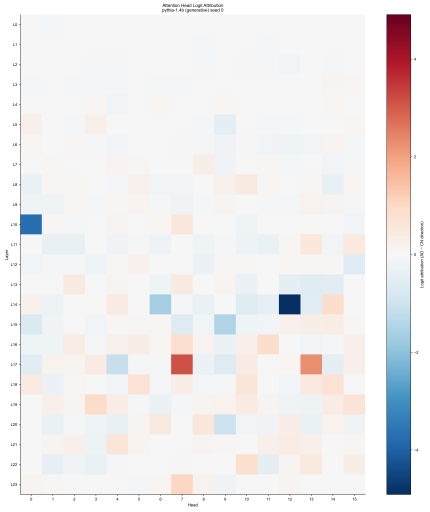
Figure 7: Logit attribution scores for Generative 160M pretrained Pythia and 12L d=256 and d=768 from scratch GPT-NeoX models at Seed 0.



(a) Classifier 1.4B pretrained Pythia model Attention head logit attribution



(b) Classifier 1.4B pretrained Pythia model Attention head logit attribution

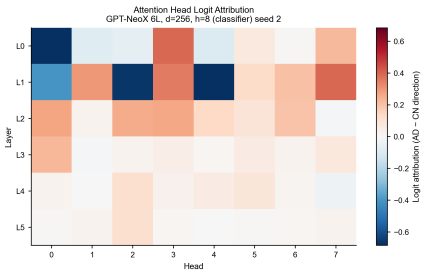


(c) Generative 1.4B pretrained Pythia model Attention head logit attribution

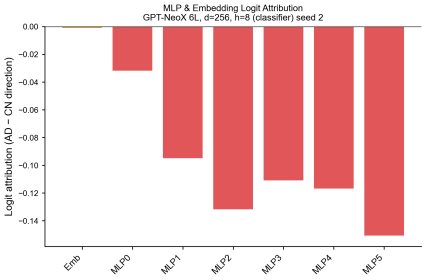


(d) Generative 1.4B pretrained Pythia model Attention head logit attribution

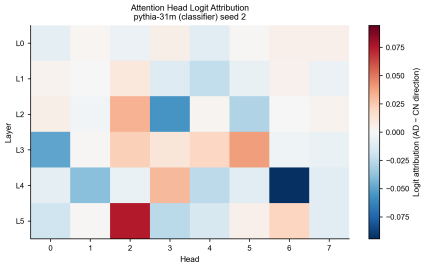
Figure 8: Logit attribution scores for Classifier and Generative 1.4B pretrained Pythia at Seed 0.



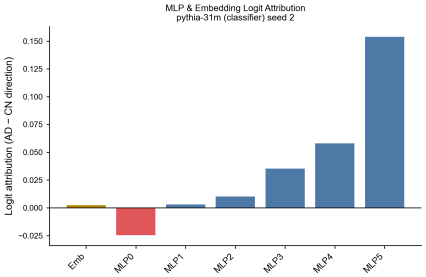
(a) Classifier 6-layer model Attention head logit attribution



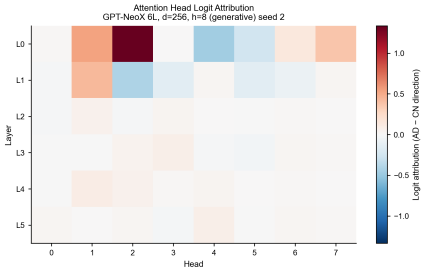
(b) Classifier 6-layer GPT-NeoX model MLP logit attribution



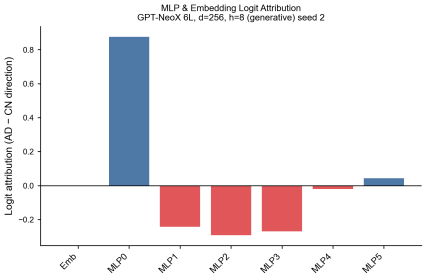
(c) Classifier 31M pretrained Pythia model Attention head logit attribution



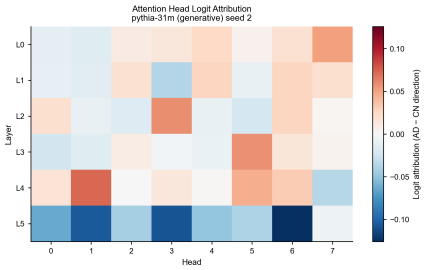
(d) Classifier 31M pretrained Pythia model Attention head logit attribution



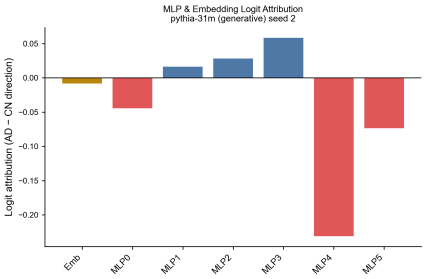
(e) Generative 6-layer model Attention head logit attribution



(f) Generative 6-layer GPT-NeoX model MLP logit attribution

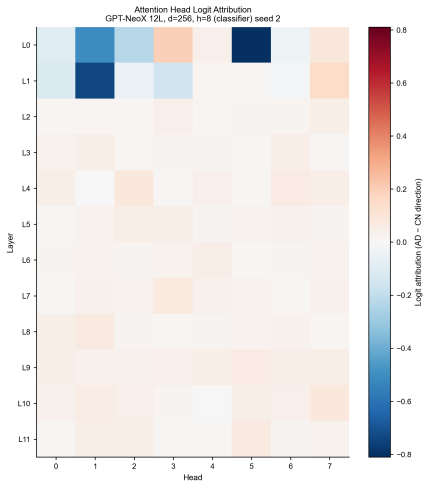


(g) Generative 31M pretrained Pythia model Attention head logit attribution

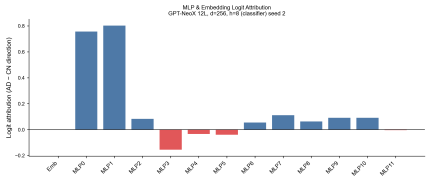


(h) Generative 31M pretrained Pythia model Attention head logit attribution

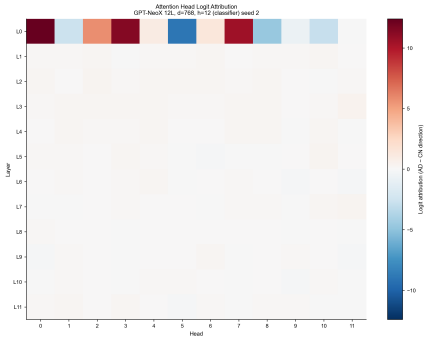
Figure 9: Logit attribution scores for Classifier and Generative 31M pretrained Pythia and 6L from-scratch GPT-NeoX Models at Seed 2.



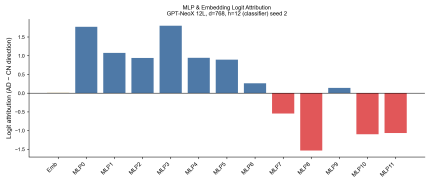
(a) Classifier 12-layer d=256 GPT-NeoX model Attention head logit attribution



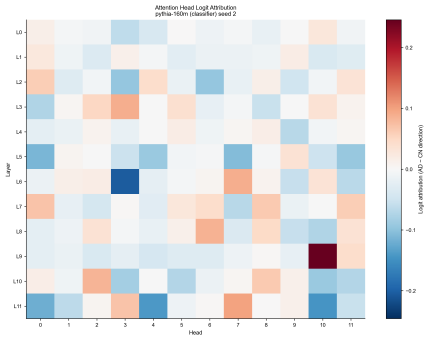
(b) Classifier 12-layer d=256 GPT-NeoX model MLP logit attribution



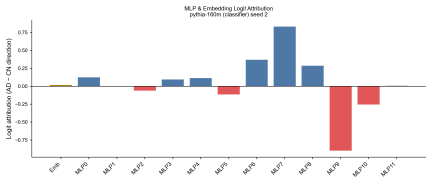
(c) Classifier 12-layer d=768 GPT-NeoX model Attention head logit attribution



(d) Classifier 12-layer d=768 GPT-NeoX model Attention head logit attribution

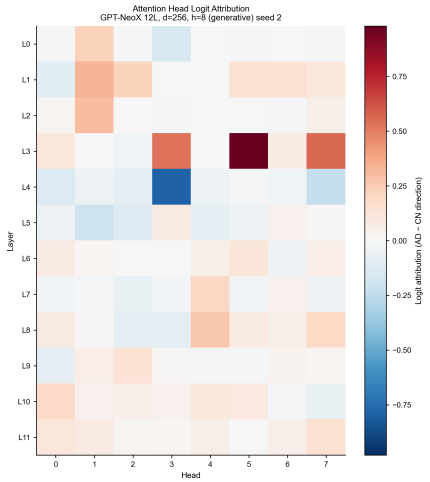


(e) Classifier 160M pretrained Pythia model Attention head logit attribution

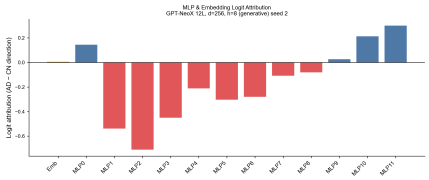


(f) Classifier 160M pretrained Pythia model Attention head logit attribution

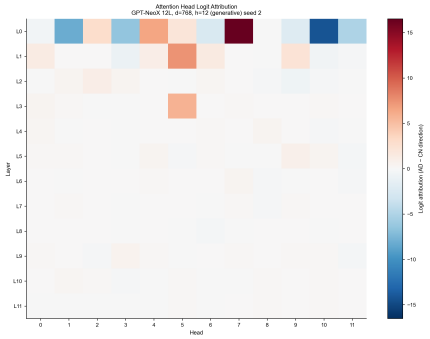
Figure 10: Logit attribution scores for Classifier 160M pretrained Pythia and 12L d=256 and d=768 from scratch GPT-NeoX models at Seed 0.



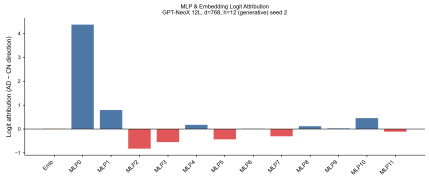
(a) Generative 12-layer d=256 GPT-NeoX model Attention head logit attribution



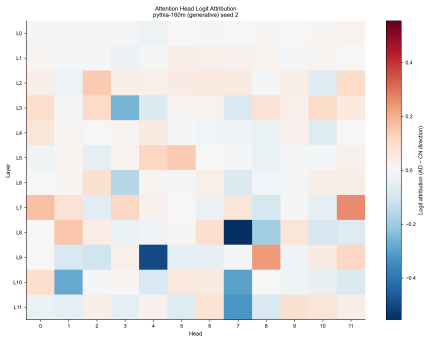
(b) Generative 12-layer d=256 GPT-NeoX model MLP logit attribution



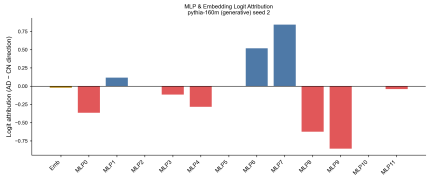
(c) Generative 12-layer d=768 GPT-NeoX model Attention head logit attribution



(d) Generative 12-layer d=768 GPT-NeoX model Attention head logit attribution

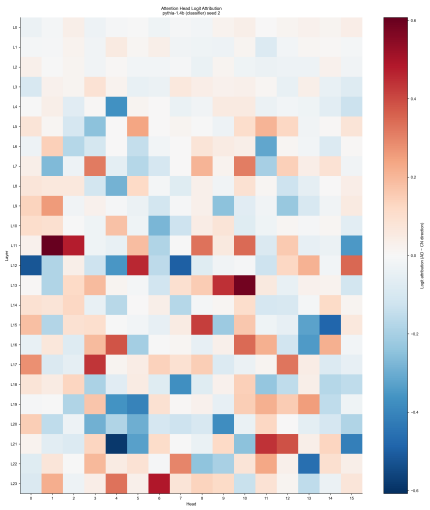


(e) Generative 160M pretrained Pythia model Attention head logit attribution



(f) Generative 160M pretrained Pythia model Attention head logit attribution

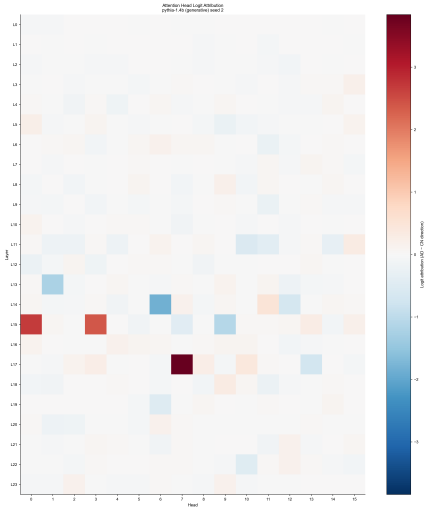
Figure 11: Logit attribution scores for Generative 160M pretrained Pythia and 12L d=256 and d=768 from scratch GPT-NeoX models at Seed 2.



(a) Classifier 1.4B pretrained Pythia model Attention head logit attribution



(b) Classifier 1.4B pretrained Pythia model Attention head logit attribution



(c) Generative 1.4B pretrained Pythia model Attention head logit attribution



(d) Generative 1.4B pretrained Pythia model Attention head logit attribution

Figure 12: Logit attribution scores for Classifier and Generative 1.4B pretrained Pythia at Seed 2.

C.3 Attention-pattern visualisation

Attention maps are generated for heads highlighted by both direct logit attribution and activation patching. These visualisations are used to inspect which positions a high-impact head reads from at the task read-out. Attention weights are reported only as descriptive diagnostics and are not interpreted as causal importance without a corresponding ablation or patching effect.

A head is reported only if it is selected independently by attribution (Appendix C.2.1 and by patching (Appendix D.1.1), under the same rule, for the same class direction.

Selection rule. Within each direction, heads are ranked by their share of that direction's total attention-head attribution; the three strongest are retained, extended to at most five while the next head carries at least half the strongest head's share. Shares are used because fixed top-k ignores concentration, a fixed share threshold returns empty sets on diffuse models, and raw logit magnitudes are not comparable across scales (per-head contributions span 0.07–16 logits).

Patching gates. Patching applies two gates in sequence. Each head's per-pair effects E are first screened by a two-sided Wilcoxon signed-rank test over usable pairs, Benjamini–Hochberg corrected across the $n_{\text{layers}} \times n_{\text{heads}}$ grid at $q \leq 0.05$. The screen establishes only that a head's effect is distinguishable from zero; it cannot rank heads by importance. With several hundred pairs per model, the test has sufficient power to detect median effects as small as $\sim 10^{-3}$, so statistical significance is reached by nearly every head regardless of practical relevance — in the 6-layer classifier, 47 of 48 heads pass the screen. Selection therefore requires the effect-size rule: screened heads are ranked by $|\text{median } E|$ and the top-three, up-to-five, half-the-strongest rule is applied. The absolute value retains heads whose patches move the recipient diagnosis significantly irrespective of the direction. Directionality pairing of the filtered list reconciles this.

Pairing and interpretation. Patches promoting AD are compared against the AD-directed attribution shortlist, and conversely. Extracted circuit membership is reported but not used for selection, since the circuits are insufficiently discriminating at scale (358 of 384 heads at 1.4B). Two constraints bound the result. Patching measures direct effects at the read-out position; a head writing at an earlier position for downstream use is invisible here. And significant single-head effects are an order of magnitude smaller than full-layer patches (~ 0.09 versus ~ 0.45), so joint implication identifies participating heads, not individually sufficient ones.

C.3.1 Attention-head pattern results

For the heads preselected by logit attribution and activation patching, we visualise attention over the tokens of the generated clinical texts, highlighting only tokens within the top 30

Across models, attention concentrates on a small set of FAQ items. Attention heads in the from-scratch models focus on clinically relevant features: heads in the 6-layer classifier attend to preparing taxes, remembering appointments, and keeping track of current events, with the 6-layer generative model additionally attending to travelling by public transport. The 12-layer classifier heads cover all of these, further extending to playing games and hobbies, paying attention to television programmes, and heating water. Whereas the 12-layer classifier attends to several functional features, the 12-layer generative model's attention is narrowly focused on the shopping item.

In contrast, for the pretrained-and-adapted 31M model, the attention heads of both the classifier and generative variants focus exclusively on numerical features, namely the FAQ total and the severity score for remembering appointments. The 160M generative model focuses narrowly on the television-attention item, and the 1.4B classifier, which is sensitive to spurious, clinically invariant wording differences and token length, attends to the numerical intrusion scores, which in isolation are not particularly clinically informative.

Table 12: Attention heads implicated by *both* read-out activation patching and logit attribution. Patching is causal and per pair: one head's read-out activation is replaced by the matched partner's, and $E = (\text{patched} - \text{recipient})/(\text{donor} - \text{recipient})$ is 0 if nothing moved and 1 if the recipient was fully converted; a NEGATIVE median means that head moved the recipient *away* from the donor's diagnosis, which is a causal involvement too and is why heads are ranked on $|E|$. A head enters the patching set only if it clears a Benjamini–Hochberg screen at $q \leq 0.05$ over the whole $n_\ell \times n_h$ grid (Wilcoxon signed-rank over the usable pairs, two-sided) *and* then lands in a share-based shortlist – the 3 strongest, extended to at most 5 while a head still carries 0.5 of the strongest. The screen alone is not a criterion: with a few hundred pairs a median effect of 0.001 is detectable, and on the 6L/d256 classifier 47 of 48 heads clear it, so an intersection against that set would return the attribution shortlist unchanged. The attribution side uses the identical selection rule, so the two are chosen symmetrically. *Attr. share* is the head's share of its direction's total attribution; *Patch median* is the median of E over pairs with a percentile-bootstrap 95% interval. Significant is not the same as large: a single head moves the read-out by ~ 0.09 where a whole attention layer moves it by ~ 0.45. AD→CN patching converts a CN recipient toward AD and is therefore compared against the AD-ward attribution heads, and conversely. *In circuit* records whether stage-2 circuit extraction also retained the head; it is reported, not used to select, because at 1.4B that circuit keeps 358 of 384 heads. Head effects here are *direct* only – a head writing at an earlier position for a later layer to read is invisible to a read-out-position patch. 32 of 147 shortlisted attribution heads across 20 model-seed pairs are also implicated causally; the accompanying CSV carries the per-model set sizes and a hypergeometric p for each overlap against chance.

Condition	Seed	Dir.	Head	Attr. share	Patch median	95% CI	In circuit
Pretrained and LoRA-adapted Pythia							
31M clf	0	CN→AD	L4H3	15%	0.623	[0.609, 0.644]	yes
31M clf	2	AD→CN	L5H2	24%	0.319	[0.303, 0.331]	yes
31M gen	0	AD→CN	L4H5	18%	0.175	[0.160, 0.181]	yes
31M gen	2	AD→CN	L4H5	8%	0.813	[0.798, 0.828]	yes
160M clf	2	CN→AD	L5H0	3%	0.176	[0.170, 0.184]	yes
160M gen	0	CN→AD	L11H7	6%	0.173	[0.150, 0.192]	–
160M gen	2	CN→AD	L11H7	7%	0.246	[0.225, 0.265]	yes
160M gen	2	CN→AD	L10H1	6%	0.220	[0.201, 0.242]	yes
1.4B clf	0	AD→CN	L9H10	2%	0.073	[0.068, 0.077]	yes
1.4B clf	0	CN→AD	L11H5	5%	0.065	[0.063, 0.068]	yes
1.4B clf	2	AD→CN	L11H2	2%	−0.080	[−0.084, −0.074]	yes
1.4B gen	0	CN→AD	L14H6	4%	0.109	[0.107, 0.111]	yes
1.4B gen	2	CN→AD	L14H6	10%	0.122	[0.119, 0.128]	yes
Architecture-matched from-scratch GPT-NeoX controls							
6L/d256 clf	0	AD→CN	L0H6	18%	0.080	[0.073, 0.088]	yes
6L/d256 clf	0	CN→AD	L1H4	16%	0.069	[0.066, 0.073]	yes
6L/d256 clf	2	CN→AD	L1H4	25%	0.109	[0.102, 0.116]	yes
6L/d256 clf	2	CN→AD	L1H2	24%	0.150	[0.135, 0.161]	yes
6L/d256 gen	0	AD→CN	L0H7	24%	0.146	[0.139, 0.152]	yes
6L/d256 gen	0	AD→CN	L0H0	22%	0.148	[0.141, 0.158]	yes
6L/d256 gen	0	CN→AD	L0H4	25%	0.084	[0.080, 0.090]	yes
6L/d256 gen	2	AD→CN	L0H2	38%	0.147	[0.144, 0.152]	yes
6L/d256 gen	2	CN→AD	L0H4	28%	0.214	[0.206, 0.218]	yes
6L/d256 gen	2	CN→AD	L1H2	25%	0.250	[0.241, 0.259]	yes
6L/d256 gen	2	CN→AD	L0H5	16%	0.161	[0.156, 0.167]	yes
12L/d768 clf	0	AD→CN	L0H8	18%	0.080	[0.073, 0.088]	yes
12L/d768 clf	0	CN→AD	L0H11	16%	0.081	[0.077, 0.087]	yes
12L/d768 clf	0	CN→AD	L0H9	15%	0.094	[0.081, 0.104]	yes
12L/d768 clf	0	CN→AD	L0H5	9%	0.117	[0.111, 0.124]	yes
12L/d768 clf	2	AD→CN	L0H3	20%	0.288	[0.107, 0.310]	yes
12L/d768 clf	2	CN→AD	L0H8	22%	0.226	[0.218, 0.246]	yes
12L/d768 gen	2	CN→AD	L0H1	18%	0.221	[0.109, 0.230]	yes
12L/d768 gen	2	CN→AD	L0H3	14%	0.119	[0.114, 0.126]	yes

C.4 LAYER-WISE PROBING OF DIAGNOSTIC AND CLINICAL INFORMATION

Disease-label probes. For each layer, residual-stream representations are extracted both at the read-out position and by mean-pooling over all sequence positions. Three probe families are evaluated: (i) a difference-of-means probe on the read-out state, (ii) logistic regression on the read-out state, and (iii) logistic regression on the mean-pooled state. Logistic-regression inputs are standardised using training-set statistics and the regularisation parameter $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ is selected by five-fold cross-validation within the probe-training split. Probe performance is then evaluated on the held-out patient subset. These analyses test where disease-label information is linearly accessible; they do not establish that the model uses that information.

Cognitive-feature probes. Continuous cognitive and functional variables are decoded from the read-out residual stream using ridge regression with $\alpha = 1$. Patients are split 80/20 at the patient level, stratified by diagnostic label; feature-specific missing values are excluded separately for train and test. Representations are standardised on the probe-training subset, and held-out R^2 and Pearson correlation are reported per feature and layer. We additionally distinguish features that are never linearly encoded from features that are encoded at an earlier layer and no longer linearly readable at the final layer. This latter pattern is described as suppression of *linear decodability*, not proof that the underlying information has been erased.

C.4.1 PROBING RESULTS

Disease-label information is already strongly linearly decodable in the earliest layers of the from-scratch controls and varies relatively little across depth: the 6-layer models peak at Layer 0 and the 12-layer models only slightly later. This establishes early availability of diagnostic information, but not its causal use.

Table 13: Linear AD-vs-CN probes on the residual stream of each from-scratch GPT-NeoX control (seeds 0 and 2, the two seeds of one model on consecutive rows). Three probes per layer on one 80/20 patient split of the IID_eor test narratives: a difference-of-means direction on the last-token residual with the threshold at the midpoint (no fitted parameters), logistic regression on the last-token residual, and logistic regression on the residual mean-pooled over all positions; the regularisation strength is chosen by 5-fold cross-validation on the training 80%. *Peak acc.* is the largest held-out accuracy over layers and *Layer* where it occurs; *Final-layer acc.* is the accuracy at the last layer. Held-out $n = 229$, so the granularity is 0.44 percentage points. *Null* p_{95} and p come from a permuted-label null with $B = 500$ replicates: the AD/CN labels are permuted across patients, the same 80/20 split and the same probe are refitted, and the SAME maximum-over-layers is taken, so the null measures the upward bias of choosing the best layer rather than the fit alone. p is one-sided, $(1 + \#\{\text{null} \geq \text{observed}\})/(1 + B)$, with floor $1/501$. Within the null the regularisation strength is held at the value the true-label fit selected for that layer: the null calibrates the choice of layer, with the probe family held fixed. Probing was performed on the from-scratch controls only.

Model	Formulation	Probe	Seed	Peak acc.	Layer	Null p_{95}	p	Final-layer acc.
6L, $d = 256$, $h = 8$	classifier	difference of means	0	79.5%	5	56.3%	<0.002	79.5%
6L, $d = 256$, $h = 8$	classifier	difference of means	2	78.6%	2	56.3%	<0.002	76.4%
6L, $d = 256$, $h = 8$	classifier	LR, last token	0	80.3%	3	57.2%	<0.002	79.0%
6L, $d = 256$, $h = 8$	classifier	LR, last token	2	79.0%	4	57.6%	<0.002	78.6%
6L, $d = 256$, $h = 8$	classifier	LR, mean-pooled	0	80.8%	3	57.6%	<0.002	80.3%
6L, $d = 256$, $h = 8$	classifier	LR, mean-pooled	2	79.9%	0	56.8%	<0.002	76.0%
6L, $d = 256$, $h = 8$	generative	difference of means	0	79.0%	0	56.3%	<0.002	76.9%
6L, $d = 256$, $h = 8$	generative	difference of means	2	79.5%	1	55.9%	<0.002	77.7%
6L, $d = 256$, $h = 8$	generative	LR, last token	0	79.0%	0	56.8%	<0.002	77.3%
6L, $d = 256$, $h = 8$	generative	LR, last token	2	79.9%	3	57.2%	<0.002	79.9%
6L, $d = 256$, $h = 8$	generative	LR, mean-pooled	0	79.0%	4	57.6%	<0.002	76.9%
6L, $d = 256$, $h = 8$	generative	LR, mean-pooled	2	79.9%	4	57.2%	<0.002	79.5%
12L, $d = 256$, $h = 8$	classifier	difference of means	0	79.9%	4	56.8%	<0.002	78.2%
12L, $d = 256$, $h = 8$	classifier	difference of means	2	79.9%	0	56.3%	<0.002	78.6%
12L, $d = 256$, $h = 8$	classifier	LR, last token	0	79.5%	2	58.1%	<0.002	79.5%
12L, $d = 256$, $h = 8$	classifier	LR, last token	2	79.5%	6	57.2%	<0.002	79.0%
12L, $d = 256$, $h = 8$	classifier	LR, mean-pooled	0	78.6%	0	57.6%	<0.002	77.3%
12L, $d = 256$, $h = 8$	classifier	LR, mean-pooled	2	79.0%	0	57.2%	<0.002	76.0%
12L, $d = 256$, $h = 8$	generative	difference of means	0	79.9%	1	56.4%	<0.002	78.2%
12L, $d = 256$, $h = 8$	generative	difference of means	2	78.6%	8	57.2%	<0.002	78.2%
12L, $d = 256$, $h = 8$	generative	LR, last token	0	80.8%	5	57.6%	<0.002	79.0%
12L, $d = 256$, $h = 8$	generative	LR, last token	2	78.6%	0	58.1%	<0.002	77.3%
12L, $d = 256$, $h = 8$	generative	LR, mean-pooled	0	79.0%	0	57.7%	<0.002	77.7%
12L, $d = 256$, $h = 8$	generative	LR, mean-pooled	2	79.9%	1	58.5%	<0.002	78.6%
12L, $d = 768$, $h = 12$	classifier	difference of means	0	77.3%	2	56.8%	<0.002	77.3%
12L, $d = 768$, $h = 12$	classifier	difference of means	2	79.5%	0	55.9%	<0.002	79.5%
12L, $d = 768$, $h = 12$	classifier	LR, last token	0	78.6%	0	57.6%	<0.002	78.2%
12L, $d = 768$, $h = 12$	classifier	LR, last token	2	79.0%	0	57.6%	<0.002	79.0%
12L, $d = 768$, $h = 12$	classifier	LR, mean-pooled	0	78.6%	0	57.6%	<0.002	76.9%
12L, $d = 768$, $h = 12$	classifier	LR, mean-pooled	2	79.9%	11	56.8%	<0.002	79.9%
12L, $d = 768$, $h = 12$	generative	difference of means	0	79.5%	0	56.3%	<0.002	78.6%
12L, $d = 768$, $h = 12$	generative	difference of means	2	78.2%	1	56.3%	<0.002	77.7%
12L, $d = 768$, $h = 12$	generative	LR, last token	0	79.5%	0	56.8%	<0.002	79.0%
12L, $d = 768$, $h = 12$	generative	LR, last token	2	80.3%	6	57.6%	<0.002	79.0%
12L, $d = 768$, $h = 12$	generative	LR, mean-pooled	0	80.3%	0	57.6%	<0.002	79.9%
12L, $d = 768$, $h = 12$	generative	LR, mean-pooled	2	78.2%	8	57.6%	<0.002	76.4%

Clinical-feature probes show strong cross-model consistency in which feature is best encoded per cognitive domain: in four of the five domains, all four from-scratch controls select the same top feature at both seeds. FAQ-derived functional measures dominate. FAQ Total is the best-encoded executive feature in every model and seed (peak held-out R^2 0.856–0.980), the FAQ item on remembering important dates leads the memory-retrieval domain (0.743–0.836), and the current-affairs item leads attention (0.462–0.625).

The one alternating domain is memory encoding, where the top feature switches between AVLT Trials 4 and 5 across models and seeds. This is expected. In the AVLT, the same word list is read to the patient five times, with recall scored after each repetition (Trial 1–5). Healthy individuals typically recall nearly the whole list by the fourth repetition, while patients with dementia continue to miss items, so Trials 4 and 5 are two nearly redundant measurements of the same underlying

ability and both separate the groups well. Which of the two a probe finds marginally easier to decode is therefore essentially arbitrary, as the two are highly correlated input features.

Clock Drawing Total is likewise consistently the better-encoded of the two visuospatial features. Conversely, several error-count features, namely intrusions, omission and commission errors, and perseverations, are not linearly decodable in any model, again with complete agreement across the set. While linear decodability does not by itself establish that a feature is used, a feature the model represents is a natural candidate for one it relies on, so the causal experiments focus on this consistently decodable set.

Table 14: Best-encoded cognitive feature per domain for each from-scratch GPT-NeoX control (seeds 0 and 2, the two seeds of one model on consecutive rows). Ridge probe on the last-token residual stream, one probe per layer and feature, scored by held-out R^2 on the 20% of patients withheld from the IID_eor test split. Peak R^2 is the maximum held-out value across layers for the single best feature in the domain, and *Layer* is where it occurs; because it is a maximum over both layers and features it is upward biased, so *Median* R^2 – the median of the per-feature peaks across the domain – and n, the number of candidate features in the domain, are given beside it: a domain carried by one outlier feature shows a large gap between the two. *Null* p_{95} and p come from a permuted-label null with $B = 1000$ replicates. Each replicate draws ONE patient permutation and applies it to every feature – the cognitive features are strongly correlated and the choice of best feature feeds on that, so permuting them independently would give a null that is too small – then refits the same ridge at every layer and takes the SAME maximum over layers and over the domain's features that *Peak* R^2 went through. The null therefore measures the upward bias of that double selection, not the fit alone. p is one-sided, $(1 + \#\{\text{null} \geq \text{observed}\})/(1 + B)$, with floor $1/1001$. Probing was performed on the from-scratch controls only.

Model/domain	Best-encoded feature	Seed	Layer	Peak R^2	Null p_{95}	p	Median R^2	n
6L d256 clf / Memory encoding	AVLT Trial4 Total	0	1	0.728	0.026	<0.001	0.699	5
6L d256 clf / Memory encoding	AVLT Trial5 Total	2	0	0.598	0.027	<0.001	0.518	5
6L d256 clf / Memory retrieval	FAQ Dates	0	2	0.832	0.025	<0.001	0.610	5
6L d256 clf / Memory retrieval	FAQ Dates	2	4	0.825	0.028	<0.001	0.472	5
6L d256 clf / Executive	FAQ Total	0	0	0.980	0.033	<0.001	0.731	10
6L d256 clf / Executive	FAQ Total	2	5	0.973	0.034	<0.001	0.733	10
6L d256 clf / Attention	FAQ CurrentAffairs	0	4	0.622	0.021	<0.001	0.179	3
6L d256 clf / Attention	FAQ CurrentAffairs	2	5	0.625	0.023	<0.001	0.229	3
6L d256 clf / Visuospatial	Clock Drawing Total	0	0	0.795	0.017	<0.001	0.480	2
6L d256 clf / Visuospatial	Clock Drawing Total	2	5	0.762	0.019	<0.001	0.455	2
6L d256 gen / Memory encoding	AVLT Trial4 Total	0	5	0.509	0.021	<0.001	0.412	5
6L d256 gen / Memory encoding	AVLT Trial4 Total	2	5	0.675	0.025	<0.001	0.568	5
6L d256 gen / Memory retrieval	FAQ Dates	0	5	0.808	0.024	<0.001	0.431	5
6L d256 gen / Memory retrieval	FAQ Dates	2	2	0.836	0.028	<0.001	0.511	5
6L d256 gen / Executive	FAQ Total	0	5	0.964	0.032	<0.001	0.706	10
6L d256 gen / Executive	FAQ Total	2	0	0.950	0.034	<0.001	0.687	10
6L d256 gen / Attention	FAQ CurrentAffairs	0	5	0.574	0.019	<0.001	0.196	3
6L d256 gen / Attention	FAQ CurrentAffairs	2	3	0.600	0.019	<0.001	0.181	3
6L d256 gen / Visuospatial	Clock Drawing Total	0	5	0.689	0.015	<0.001	0.388	2
6L d256 gen / Visuospatial	Clock Drawing Total	2	4	0.571	0.020	<0.001	0.330	2
12L d768 clf / Memory encoding	AVLT Trial5 Total	0	10	0.631	0.007	<0.001	0.584	5
12L d768 clf / Memory encoding	AVLT Trial4 Total	2	11	0.499	0.023	<0.001	0.397	5
12L d768 clf / Memory retrieval	FAQ Dates	0	6	0.807	0.010	<0.001	0.514	5
12L d768 clf / Memory retrieval	FAQ Dates	2	8	0.743	0.025	<0.001	0.426	5
12L d768 clf / Executive	FAQ Total	0	1	0.936	0.024	<0.001	0.635	10
12L d768 clf / Executive	FAQ Total	2	11	0.905	0.034	<0.001	0.642	10
12L d768 clf / Attention	FAQ CurrentAffairs	0	0	0.534	0.012	<0.001	0.113	3
12L d768 clf / Attention	FAQ CurrentAffairs	2	8	0.606	0.023	<0.001	0.106	3
12L d768 clf / Visuospatial	Clock Drawing Total	0	8	0.657	0.002	<0.001	0.372	2
12L d768 clf / Visuospatial	Clock Drawing Total	2	11	0.713	0.018	<0.001	0.413	2
12L d768 gen / Memory encoding	AVLT Trial5 Total	0	4	0.422	0.023	<0.001	0.320	5
12L d768 gen / Memory encoding	AVLT Trial5 Total	2	10	0.478	0.022	<0.001	0.433	5
12L d768 gen / Memory retrieval	FAQ Dates	0	2	0.759	0.024	<0.001	0.333	5
12L d768 gen / Memory retrieval	FAQ Dates	2	4	0.814	0.025	<0.001	0.379	5
12L d768 gen / Executive	FAQ Total	0	11	0.885	0.035	<0.001	0.624	10
12L d768 gen / Executive	FAQ Total	2	11	0.856	0.035	<0.001	0.526	10
12L d768 gen / Attention	FAQ CurrentAffairs	0	2	0.620	0.024	<0.001	0.104	3
12L d768 gen / Attention	FAQ CurrentAffairs	2	1	0.462	0.024	<0.001	0.102	3
12L d768 gen / Visuospatial	Clock Drawing Total	0	2	0.705	0.018	<0.001	0.389	2
12L d768 gen / Visuospatial	Clock Drawing Total	2	7	0.458	0.016	<0.001	0.274	2

D Causal Intervention and Counterfactual Analyses

This section corresponds to the causal-learning arm of the study. Each analysis intervenes on either an internal activation or an input feature and measures the induced change in the model's own class margin or prediction.

D.1 Layer-wise read-out activation patching

Read-out patching asks when a participant-specific diagnostic representation becomes causally effective. We form opposite-label pairs matched on sex and exact year of birth using greedy matching without replacement. On the full 1,142-example test pool the matcher yields up to 403 exact-YOB/sex pairs before model-specific gating.

For each pair, one participant is the donor and the other the recipient. At each layer we replace exactly one activation at the recipient's task-specific read-out position with the corresponding donor activation. We patch the residual stream before the block, attention output, MLP output, and individual attention-head output separately. Both AD→CN and CN→AD directions are retained rather than averaged.

If d_D, d_R, and $d_{P,\ell}$ denote donor, recipient, and patched diagnostic margins, the normalised intervention effect is

$$E_\ell = \frac{d_{P,\ell} - d_R}{d_D - d_R}. \tag{9}$$

Thus $E_\ell = 0$ leaves the recipient unchanged and $E_\ell = 1$ reaches the donor endpoint. Because this ratio becomes unstable for poorly separated donor and recipient margins, pairs are retained only when both examples are correctly classified and their baseline-margin separation is at least one model-specific standard deviation. A run with fewer than 20 usable pairs is reported as aborted. Since E_ℓ is unbounded, the primary summaries are the median and interquartile range; means and standard deviations are secondary diagnostics.

Two interpretation constraints are important. First, final-layer residual patching is a positive control because it occurs immediately upstream of the output and should approach the donor endpoint. Second, classifier and generative models use different lexical read-out positions: the classifier reads the final narrative token and the generative model reads the `Diagnosis:` cue. Patching therefore compares complete task-specific read-out dynamics rather than isolating the output head alone.

D.1.1 Read-out patching results

Read-out patching localises the causal action consistently with the attribution analysis. Final-layer residual patching approaches full donor conversion in most conditions, confirming the positive control, and effects are near-symmetric across the AD→CN and CN→AD directions throughout. The from-scratch controls are causally shallow and attention-mediated: a single attention-output swap at layer 0–1 recovers roughly half to all of the donor–recipient margin (reaching complete conversion, 1.00 $[1.00, 1.00]$, at layer 0 in the 12L/d768 wide classifier at seed 2), while MLP patches at any layer move the margin by at most 0.16. The pretrained models place the causal action mid-to-late instead: attention at layers 4–5 of 6 at 31M and layers 10–11 of 24 at 1.4B, mirroring the depth contrast seen in the direct-logit attribution. Mechanisms for pretrained models are not always stable across seeds: the 160M classifier is MLP-dominated at seed 0 (mlp L2, median $\approx$ 0.9–1.1) but attention-mediated at seed 2, the only strongly MLP-mediated condition in the set.

Two conditions fall short of the positive control. In the 31M classifier at seed 2, the final-layer residual patch recovers only about half of the donor–recipient margin (median E_ℓ of 0.47 and 0.51 in the two patching directions), and in the 160M generative model at seed 2 it recovers roughly two thirds (medians of 0.59–0.67 across seeds and directions), where all other conditions reach $\approx$0.93–1.06. Both are middle-sized models with performance instability and seed-sensitivity that the learning rate tuning tried to minimise. Therefore their patching results in these rows should be read with caution.

Table 15: Matched read-out patching. For each pair the donor's activation is spliced into the recipient at the read-out position only, and the effect is $E_\ell = (\text{patched} - \text{recipient})/(\text{donor} - \text{recipient})$: 0 leaves the recipient unchanged, 1 converts it fully to the donor's diagnosis. *Matched* is the AD–CN pairs of identical year of birth and the same sex that the test split affords; it is a property of the cohort, identical for every model. *Usable* is what survives the run's gates – both members classified correctly, and the pair separated by at least one baseline logit-difference s.d. The two columns are not interchangeable: a model can score respectably on accuracy and still leave no usable pair. Cells give the peak layer with the median [IQR] of E_ℓ over usable pairs; runs with fewer than 20 usable pairs are marked *aborted* rather than assigned an effect of zero, which applies to 0 run(s) here. Peak layers are read off the curve, not chosen by eye; the accompanying figures carry the complete layer curve. Note that `resid_pre` at the final layer transplants nearly the whole computation and approaches 1 by construction – it bounds what any intervention at this position can achieve, and is a positive control rather than a result. The informative sites are `attn_out` and `mlp_out`.

Condition	Direction	Matched	Usable	Residual control	Attn./MLP localisation
From-scratch 6L/$d256$ classifier (s0)	AD→CN	403	240	L5: 0.96 [0.95, 0.97]	attn L1: 0.45 [0.41, 0.51]; mlp L5: -0.03 [-0.03, -0.03]
From-scratch 6L/$d256$ classifier (s0)	CN→AD	403	240	L5: 0.98 [0.97, 0.98]	attn L1: 0.51 [0.43, 0.58]; mlp L1: 0.04 [0.03, 0.05]
From-scratch 6L/$d256$ classifier (s2)	AD→CN	403	232	L5: 1.00 [0.99, 1.00]	attn L1: 0.50 [0.44, 0.54]; mlp L2: 0.05 [0.04, 0.06]
From-scratch 6L/$d256$ classifier (s2)	CN→AD	403	232	L5: 0.99 [0.99, 1.00]	attn L1: 0.62 [0.55, 0.66]; mlp L2: 0.15 [0.13, 0.16]
31M classifier (s0)	AD→CN	403	239	L5: 0.93 [0.89, 0.98]	attn L4: 0.77 [0.71, 0.83]; mlp L5: 0.39 [0.35, 0.41]
31M classifier (s0)	CN→AD	403	239	L5: 0.94 [0.90, 0.98]	attn L4: 0.80 [0.75, 0.85]; mlp L5: 0.41 [0.38, 0.44]
31M classifier (s2)	AD→CN	403	260	L5: 0.47 [0.43, 0.53]	attn L5: 0.51 [0.44, 0.56]; mlp L5: 0.25 [0.22, 0.28]
31M classifier (s2)	CN→AD	403	260	L5: 0.51 [0.46, 0.57]	attn L5: 0.53 [0.48, 0.57]; mlp L5: 0.27 [0.23, 0.32]
From-scratch 6L/$d256$ generative (s0)	AD→CN	403	227	L5: 0.98 [0.98, 0.99]	attn L0: 0.66 [0.62, 0.70]; mlp L1: 0.07 [0.06, 0.08]
From-scratch 6L/$d256$ generative (s0)	CN→AD	403	227	L5: 1.00 [1.00, 1.00]	attn L0: 0.78 [0.75, 0.82]; mlp L1: 0.06 [0.05, 0.07]
From-scratch 6L/$d256$ generative (s2)	AD→CN	403	243	L5: 0.99 [0.99, 0.99]	attn L0: 0.60 [0.55, 0.64]; mlp L2: 0.03 [0.03, 0.04]
From-scratch 6L/$d256$ generative (s2)	CN→AD	403	243	L5: 0.99 [0.99, 1.00]	attn L0: 0.72 [0.67, 0.76]; mlp L1: 0.12 [0.12, 0.13]
31M generative (s0)	AD→CN	403	228	L5: 1.00 [0.97, 1.04]	attn L4: 0.85 [0.80, 0.96]; mlp L5: 0.37 [0.33, 0.40]
31M generative (s0)	CN→AD	403	228	L5: 1.00 [0.97, 1.04]	attn L4: 0.85 [0.82, 0.96]; mlp L5: 0.39 [0.34, 0.43]
31M generative (s2)	AD→CN	403	234	L5: 1.00 [0.99, 1.01]	attn L4: 0.96 [0.93, 1.00]; mlp L5: 0.15 [0.13, 0.17]
31M generative (s2)	CN→AD	403	234	L5: 1.00 [1.00, 1.01]	attn L4: 0.97 [0.93, 1.00]; mlp L5: 0.17 [0.15, 0.19]
From-scratch 12L/$d768$ classifier (s0)	AD→CN	403	230	L5: 1.01 [1.00, 1.01]	attn L0: 0.80 [0.75, 0.86]; mlp L1: 0.14 [0.13, 0.17]
From-scratch 12L/$d768$ classifier (s0)	CN→AD	403	230	L4: 1.01 [1.00, 1.03]	attn L0: 0.90 [0.84, 0.94]; mlp L1: 0.16 [0.14, 0.18]
From-scratch 12L/$d768$ classifier (s2)	AD→CN	403	252	L11: 1.00 [1.00, 1.00]	attn L0: 1.00 [1.00, 1.00]; mlp L1: 0.07 [0.06, 0.09]
From-scratch 12L/$d768$ classifier (s2)	CN→AD	403	252	L11: 1.00 [1.00, 1.00]	attn L0: 1.00 [1.00, 1.00]; mlp L1: 0.07 [0.07, 0.08]
160M classifier (s0)	AD→CN	403	180	L11: 0.95 [0.92, 0.98]	attn L1: 0.27 [0.19, 0.40]; mlp L2: 0.89 [0.72, 1.21]
160M classifier (s0)	CN→AD	403	180	L11: 0.97 [0.93, 1.00]	attn L1: 0.20 [0.15, 0.24]; mlp L2: 1.08 [0.84, 1.33]
160M classifier (s2)	AD→CN	403	162	L11: 0.97 [0.94, 0.99]	attn L5: 0.41 [0.36, 0.47]; mlp L9: 0.39 [0.35, 0.41]
160M classifier (s2)	CN→AD	403	162	L11: 0.96 [0.94, 0.99]	attn L7: 0.38 [0.33, 0.43]; mlp L9: 0.41 [0.38, 0.45]
From-scratch 12L/$d768$ generative (s0)	AD→CN	403	255	L11: 1.00 [0.99, 1.00]	attn L0: 0.76 [0.70, 0.86]; mlp L1: 0.05 [0.04, 0.06]
From-scratch 12L/$d768$ generative (s0)	CN→AD	403	255	L11: 1.00 [0.99, 1.00]	attn L0: 0.88 [0.85, 0.92]; mlp L1: 0.06 [0.06, 0.07]
From-scratch 12L/$d768$ generative (s2)	AD→CN	403	240	L11: 1.00 [1.00, 1.00]	attn L0: 0.82 [0.79, 0.86]; mlp L1: 0.07 [0.06, 0.08]
From-scratch 12L/$d768$ generative (s2)	CN→AD	403	240	L11: 1.00 [1.00, 1.00]	attn L0: 0.83 [0.79, 0.85]; mlp L1: 0.09 [0.09, 0.10]
160M generative (s0)	AD→CN	403	180	L11: 0.66 [0.55, 0.76]	attn L9: 0.44 [0.32, 0.61]; mlp L11: 0.28 [0.16, 0.51]
160M generative (s0)	CN→AD	403	180	L11: 0.67 [0.55, 0.77]	attn L9: 0.38 [0.28, 0.49]; mlp L11: 0.27 [0.16, 0.51]
160M generative (s2)	AD→CN	403	122	L11: 0.59 [0.53, 0.70]	attn L11: 0.39 [0.29, 0.47]; mlp L11: 0.15 [0.04, 0.24]
160M generative (s2)	CN→AD	403	122	L11: 0.60 [0.51, 0.71]	attn L11: 0.39 [0.29, 0.47]; mlp L11: 0.15 [0.04, 0.23]
1.4B classifier (s0)	AD→CN	403	217	L23: 0.98 [0.97, 0.99]	attn L10: 0.24 [0.20, 0.27]; mlp L23: 0.19 [0.17, 0.21]
1.4B classifier (s0)	CN→AD	403	217	L23: 0.98 [0.96, 0.99]	attn L10: 0.25 [0.22, 0.28]; mlp L23: 0.29 [0.27, 0.31]
1.4B classifier (s2)	AD→CN	403	233	L16: 1.06 [1.00, 1.11]	attn L10: 0.40 [0.35, 0.44]; mlp L22: 0.17 [0.15, 0.19]
1.4B classifier (s2)	CN→AD	403	233	L19: 1.00 [0.99, 1.03]	attn L10: 0.31 [0.14, 0.46]; mlp L22: 0.19 [0.17, 0.22]
1.4B generative (s0)	AD→CN	403	187	L16: 1.00 [0.99, 1.01]	attn L10: 0.63 [0.58, 0.68]; mlp L16: 0.07 [0.07, 0.07]
1.4B generative (s0)	CN→AD	403	187	L17: 1.00 [1.00, 1.01]	attn L10: 0.55 [0.49, 0.62]; mlp L23: 0.07 [0.06, 0.08]
1.4B generative (s2)	AD→CN	403	217	L23: 1.00 [1.00, 1.00]	attn L11: 0.46 [0.43, 0.48]; mlp L21: -0.08 [-0.10, -0.07]
1.4B generative (s2)	CN→AD	403	217	L23: 1.00 [1.00, 1.00]	attn L11: 0.70 [0.65, 0.73]; mlp L23: 0.12 [0.10, 0.14]

D.2 DEMOGRAPHIC ACTIVATION SWAPPING

Demographic swapping tests whether internal representations associated with sex, recorded race, or age-related text causally move the diagnostic decision. Opposite-label patients are greedily matched on a standardised cognitive and functional profile, subject to a profile-distance caliper of 0.6 cohort standard deviations and additional demographic constraints specific to the contrast. Sex swapping holds recorded race and year of birth approximately fixed; race swapping holds sex and year of birth approximately fixed; age swapping holds sex and race fixed and evaluates pre-specified year-of-birth gap bands.

At each layer the residual-stream activation at the target demographic span is replaced with the matched partner's activation. We report change in diagnostic margin and label-flip rate in each direction. We employ two positive control arms: an extreme opposite-label FAQ control (the activation of the most extreme FAQ total score from the opposite labelled cohort is transplanted); and a final-residual patching intervention. Mean margin changes are accompanied by bootstrap confidence intervals. These experiments are not a fairness audit and do not justify claims about absence of demographic bias.

Another limitation is that the cohort that results from cognitive matching systematically selects for the mildest forms of AD, and excludes severe forms. Therefore, we may not see the true causal effect of some demographic features as cognitively matching is inherently ill-posed. We minimise this effect by doing the extreme FAQ donor arm, which transplants 0 (the lowest possible score for CN), and 30 (the highest possible score for AD) in opposite labeled clinical summaries.

D.2.1 DEMOGRAPHIC-SWAPPING RESULTS

Demographic swaps are near-null throughout. Across all four controls, both formulations, both seeds, and all attribute contrasts, transplanting the partner's residual activations at the demographic token span moves the diagnostic margin by at most $|\Delta| \approx 0.08$ logits, and label-flip rates never exceed 5% (most are zero, with Wilson intervals bounded by small n rather than by observed flips). The interventions themselves are verified: the final-residual positive control behaves as expected in every run.

The clearest departures from zero are small age effects in the 6L model: in the classifier, and in the generative model at seed 0, the wider year-of-birth bands produce peak margin changes of $+0.02$ to $+0.08$ with bootstrap intervals excluding zero, growing with the size of the age contrast relative to the near-zero-contrast reference band. This suggests a weak age sensitivity. This effect is not stable, however: it is absent in the 6L generative model at seed 2 and in both 12L architectures, where age swaps are indistinguishable from zero. A single seed-specific race effect appears in the 6L classifier at seed 2 ($+0.05$ $[0.03, 0.07]$, on only 59 pairs) with no counterpart at seed 0 or in any other model.

The extreme-FAQ control produces large margin shifts and flip rates up to 0.90 in the 12L/d768 generative model. However, the extreme-FAQ control is near-null in most other runs. We clinically expect FAQ total to be a highly salient feature and did not expect such under-reliance. We explore in the next section this phenomenon, where we compare verbal as opposed to numeric interventions on cognitive scores, and find that models rely more on verbal than numerical features.

Table 16: Demographic activation swapping in the from-scratch GPT-NeoX controls (seeds 0 and 2, the two seeds of one model on consecutive rows), AD→CN (normalising edits, patients called AD). AD–CN pairs are matched on a z-scored ten-variable cognitive profile (caliper 0.6 SD) with the stated attribute contrasted and the others held the same. The residual vector at the attribute's token is transplanted from AD to CN

Condition	Attribute	n pairs	n gated	Peak Δmargin [95% CI]	Max flip [95% CI]	Layer	Extreme ctrl Δ / flip	Readout
6L/d256 classifier (s0)	sex	179	83	+0.000 [-0.001, 0.002]	**0.012** [0.002, 0.065]	0	-0.010 / 0.00	1.00
6L/d256 classifier (s2)	sex	179	87	+0.003 [-0.003, 0.009]	0.000 [0.000, 0.042]	1	-0.010 / 0.00	1.00
6L/d256 classifier (s0)	race	59	26	+0.001 [0.000, 0.001]	0.000 [0.000, 0.129]	0	-0.014 / 0.00	1.00
6L/d256 classifier (s2)	race	59	29	-0.034 [-0.061, -0.006]	0.000 [0.000, 0.117]	1	-0.007 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	83	+0.008 [-0.008, 0.026]	**0.012** [0.002, 0.065]	1	-0.011 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	94	+0.003 [-0.019, 0.024]	0.000 [0.000, 0.039]	1	-0.010 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	108	+0.008 [-0.008, 0.025]	**0.019** [0.005, 0.065]	1	-0.011 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	112	-0.002 [-0.019, 0.017]	**0.009** [0.002, 0.049]	0	-0.010 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	97	+0.007 [-0.011, 0.025]	**0.021** [0.006, 0.072]	1	-0.010 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	97	+0.005 [-0.015, 0.025]	0.000 [0.000, 0.038]	1	-0.009 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	84	-0.002 [-0.009, 0.004]	**0.024** [0.007, 0.083]	2	-0.011 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	81	+0.010 [-0.011, 0.030]	0.000 [0.000, 0.045]	1	-0.008 / 0.00	1.00
6L/d256 generative (s0)	sex	179	89	+0.001 [-0.002, 0.003]	0.000 [0.000, 0.041]	0	-0.003 / 0.00	1.00
6L/d256 generative (s2)	sex	179	89	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	0	-0.406 / **0.02**	1.00
6L/d256 generative (s0)	race	59	32	+0.003 [0.000, 0.005]	0.000 [0.000, 0.107]	0	-0.003 / 0.00	1.00
6L/d256 generative (s2)	race	59	32	+0.002 [0.001, 0.002]	0.000 [0.000, 0.107]	0	-0.400 / **0.03**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	94	+0.021 [-0.005, 0.052]	**0.011** [0.002, 0.058]	0	-0.003 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	100	-0.000 [-0.001, 0.000]	**0.010** [0.002, 0.054]	0	-0.419 / **0.03**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	113	+0.007 [-0.003, 0.019]	**0.009** [0.002, 0.048]	1	-0.003 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	118	+0.001 [-0.000, 0.003]	**0.008** [0.001, 0.046]	1	-0.427 / **0.02**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	100	+0.021 [-0.005, 0.052]	**0.010** [0.002, 0.054]	0	-0.003 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	+0.001 [-0.001, 0.003]	0.000 [0.000, 0.036]	1	-0.416 / **0.02**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	+0.017 [-0.013, 0.053]	**0.011** [0.002, 0.062]	0	-0.003 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	90	+0.001 [-0.001, 0.003]	**0.011** [0.002, 0.060]	1	-0.416 / **0.03**	1.00
12L/d256 classifier (s0)	sex	179	90	+0.001 [-0.002, 0.006]	0.000 [0.000, 0.041]	0	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	sex	179	90	+0.000 [-0.001, 0.001]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d256 classifier (s0)	race	59	32	+0.005 [0.002, 0.007]	0.000 [0.000, 0.107]	0	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	race	59	31	-0.000 [-0.000, -0.000]	0.000 [0.000, 0.110]	0	-0.000 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	95	+0.003 [-0.004, 0.016]	0.000 [0.000, 0.039]	1	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	98	+0.001 [-0.000, 0.003]	0.000 [0.000, 0.038]	0	-0.000 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	112	+0.007 [-0.010, 0.029]	**0.009** [0.002, 0.049]	1	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	115	+0.001 [-0.001, 0.002]	0.000 [0.000, 0.032]	1	-0.000 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	99	+0.009 [-0.010, 0.033]	**0.010** [0.002, 0.055]	1	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	102	+0.000 [-0.002, 0.003]	0.000 [0.000, 0.036]	0	-0.000 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	-0.003 [-0.012, 0.004]	**0.011** [0.002, 0.062]	2	+0.000 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	89	+0.002 [0.000, 0.005]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d256 generative (s0)	sex	179	93	+0.000 [-0.000, 0.001]	0.000 [0.000, 0.040]	3	-0.001 / 0.00	1.00
12L/d256 generative (s2)	sex	179	84	+0.000 [0.000, 0.000]	0.000 [0.000, 0.044]	2	-0.017 / 0.00	1.00
12L/d256 generative (s0)	race	59	32	+0.006 [0.003, 0.009]	0.000 [0.000, 0.107]	0	-0.001 / 0.00	1.00
12L/d256 generative (s2)	race	59	30	-0.009 [-0.018, -0.000]	0.000 [0.000, 0.114]	1	-0.010 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	100	+0.001 [-0.004, 0.007]	0.000 [0.000, 0.037]	0	-0.001 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	88	+0.002 [-0.003, 0.009]	0.000 [0.000, 0.042]	5	-0.017 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	118	-0.004 [-0.008, 0.001]	**0.017** [0.005, 0.060]	0	-0.001 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	109	-0.002 [-0.006, 0.003]	0.000 [0.000, 0.034]	0	-0.018 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	103	-0.003 [-0.007, 0.001]	**0.010** [0.002, 0.053]	1	-0.001 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	94	+0.001 [-0.004, 0.006]	0.000 [0.000, 0.039]	3	-0.017 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	-0.004 [-0.009, -0.000]	**0.011** [0.002, 0.062]	1	-0.001 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	76	-0.001 [-0.004, 0.003]	0.000 [0.000, 0.048]	2	-0.014 / 0.00	1.00
12L/d768 classifier (s0)	sex	179	87	+0.001 [-0.000, 0.003]	0.000 [0.000, 0.042]	0	-0.006 / 0.00	1.00
12L/d768 classifier (s2)	sex	179	91	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	race	59	33	+0.008 [-0.000, 0.022]	0.000 [0.000, 0.104]	0	+0.000 / 0.00	1.00
12L/d768 classifier (s2)	race	59	34	+0.001 [0.000, 0.001]	0.000 [0.000, 0.102]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	89	+0.000 [0.000, 0.001]	0.000 [0.000, 0.041]	0	-0.004 / 0.00	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	101	+0.000 [-0.000, 0.002]	0.000 [0.000, 0.037]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	119	+0.000 [-0.000, 0.001]	0.000 [0.000, 0.031]	0	-0.004 / 0.00	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	119	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.031]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	+0.000 [-0.000, 0.001]	0.000 [0.000, 0.036]	0	-0.005 / 0.00	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.036]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	+0.000 [-0.000, 0.001]	0.000 [0.000, 0.042]	0	-0.006 / 0.00	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	91	-0.000 [-0.001, 0.000]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d768 generative (s0)	sex	179	91	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	2	-0.996 / **0.40**	1.00
12L/d768 generative (s2)	sex	179	84	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.044]	0	+0.001 / 0.00	1.00
12L/d768 generative (s0)	race	59	33	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.104]	0	-1.056 / **0.53**	1.00
12L/d768 generative (s2)	race	59	31	-0.001 [-0.001, -0.000]	0.000 [0.000, 0.110]	0	+0.001 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	98	-0.003 [-0.006, -0.001]	0.000 [0.000, 0.038]	1	-1.011 / **0.35**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	88	+0.001 [-0.000, 0.003]	0.000 [0.000, 0.042]	1	+0.002 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	117	-0.002 [-0.005, 0.001]	0.000 [0.000, 0.032]	0	-1.032 / **0.36**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	110	-0.001 [-0.004, 0.002]	0.000 [0.000, 0.034]	0	+0.002 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	103	-0.001 [-0.005, 0.003]	0.000 [0.000, 0.036]	0	-1.014 / **0.38**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	96	-0.001 [-0.004, 0.002]	0.000 [0.000, 0.038]	0	+0.002 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	90	-0.002 [-0.006, 0.001]	0.000 [0.000, 0.041]	0	-1.026 / **0.39**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	-0.002 [-0.006, 0.002]	0.000 [0.000, 0.042]	0	+0.002 / 0.00	1.00

Table 17: Demographic activation swapping in the from-scratch GPT-NeoX controls, CN→AD (impairing edits, patients called CN). AD–CN pairs are matched on a z-scored ten-variable cognitive profile (caliper 0.6 SD) with the stated attribute contrasted and the others held the same. The residual vector at the attribute's token is transplanted from CN to AD

Condition	Attribute	n pairs	n gated	Peak Δmargin [95% CI]	Max flip [95% CI]	Layer	Extreme ctrl Δ / flip	Readout
6L/d256 classifier (s0)	sex	179	83	-0.001 [-0.003, 0.000]	0.000 [0.000, 0.044]	0	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	sex	179	87	-0.005 [-0.015, 0.005]	**0.011** [0.002, 0.062]	0	-0.004 / 0.00	1.00
6L/d256 classifier (s0)	race	59	26	-0.001 [-0.001, -0.000]	0.000 [0.000, 0.129]	0	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	race	59	29	+0.046 [0.026, 0.068]	0.000 [0.000, 0.117]	1	-0.003 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	83	+0.016 [-0.005, 0.038]	**0.024** [0.007, 0.084]	1	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	94	+0.004 [-0.015, 0.023]	0.000 [0.000, 0.039]	1	-0.005 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	108	+0.029 [0.002, 0.057]	**0.019** [0.005, 0.065]	1	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	112	+0.023 [0.005, 0.039]	0.000 [0.000, 0.033]	1	-0.004 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	97	+0.037 [0.007, 0.069]	**0.031** [0.011, 0.087]	1	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	97	+0.021 [0.001, 0.040]	0.000 [0.000, 0.038]	1	-0.004 / 0.00	1.00
6L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	84	+0.031 [0.002, 0.062]	**0.024** [0.007, 0.083]	1	-0.002 / 0.00	1.00
6L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	81	+0.026 [0.006, 0.048]	0.000 [0.000, 0.045]	1	-0.003 / 0.00	1.00
6L/d256 generative (s0)	sex	179	89	-0.001 [-0.004, 0.003]	0.000 [0.000, 0.041]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	sex	179	89	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	1	+0.185 / **0.07**	1.00
6L/d256 generative (s0)	race	59	32	-0.008 [-0.011, -0.004]	0.000 [0.000, 0.107]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	race	59	32	-0.002 [-0.004, -0.001]	0.000 [0.000, 0.107]	0	+0.185 / **0.06**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	94	+0.028 [-0.012, 0.068]	**0.032** [0.011, 0.090]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	100	+0.001 [-0.000, 0.003]	0.000 [0.000, 0.037]	0	+0.186 / **0.09**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	113	+0.060 [0.010, 0.113]	**0.027** [0.009, 0.075]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	118	+0.000 [-0.003, 0.003]	0.000 [0.000, 0.032]	0	+0.186 / **0.08**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	100	+0.061 [0.006, 0.116]	**0.040** [0.016, 0.098]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	+0.001 [-0.003, 0.004]	0.000 [0.000, 0.036]	0	+0.194 / **0.11**	1.00
6L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	+0.076 [0.017, 0.142]	**0.045** [0.018, 0.111]	0	+0.001 / 0.00	1.00
6L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	90	+0.002 [-0.003, 0.006]	0.000 [0.000, 0.041]	0	+0.207 / **0.09**	1.00
12L/d256 classifier (s0)	sex	179	90	-0.002 [-0.004, 0.000]	0.000 [0.000, 0.041]	0	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	sex	179	90	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	0	-0.001 / 0.00	1.00
12L/d256 classifier (s0)	race	59	32	-0.002 [-0.003, -0.001]	0.000 [0.000, 0.107]	0	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	race	59	31	-0.000 [-0.000, -0.000]	0.000 [0.000, 0.110]	1	-0.001 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	95	-0.005 [-0.018, 0.003]	0.000 [0.000, 0.039]	2	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	98	+0.001 [-0.000, 0.003]	0.000 [0.000, 0.038]	1	-0.001 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	112	-0.006 [-0.021, 0.005]	**0.009** [0.002, 0.049]	2	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	115	+0.002 [-0.000, 0.004]	0.000 [0.000, 0.032]	0	-0.001 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	99	-0.006 [-0.023, 0.006]	**0.010** [0.002, 0.055]	2	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	102	+0.002 [-0.000, 0.004]	0.000 [0.000, 0.036]	0	-0.001 / 0.00	1.00
12L/d256 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	-0.007 [-0.026, 0.007]	**0.011** [0.002, 0.062]	2	-0.001 / 0.00	1.00
12L/d256 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	89	+0.002 [-0.000, 0.005]	0.000 [0.000, 0.041]	1	-0.001 / 0.00	1.00
12L/d256 generative (s0)	sex	179	93	-0.001 [-0.002, 0.000]	0.000 [0.000, 0.040]	0	+0.000 / 0.00	1.00
12L/d256 generative (s2)	sex	179	84	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.044]	0	-0.007 / 0.00	1.00
12L/d256 generative (s0)	race	59	32	-0.009 [-0.014, -0.004]	0.000 [0.000, 0.107]	0	-0.000 / 0.00	1.00
12L/d256 generative (s2)	race	59	30	-0.009 [-0.023, 0.004]	0.000 [0.000, 0.114]	0	-0.007 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	100	-0.003 [-0.012, 0.005]	0.000 [0.000, 0.037]	0	+0.000 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	88	+0.001 [-0.006, 0.009]	0.000 [0.000, 0.042]	1	-0.007 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	118	+0.003 [-0.005, 0.011]	0.000 [0.000, 0.032]	1	+0.000 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	109	+0.003 [-0.003, 0.010]	0.000 [0.000, 0.034]	1	-0.007 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	103	-0.007 [-0.024, 0.009]	0.000 [0.000, 0.036]	0	+0.000 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	94	-0.001 [-0.005, 0.002]	0.000 [0.000, 0.039]	4	-0.007 / 0.00	1.00
12L/d256 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	-0.007 [-0.026, 0.012]	0.000 [0.000, 0.042]	0	+0.000 / 0.00	1.00
12L/d256 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	76	+0.001 [-0.003, 0.006]	0.000 [0.000, 0.048]	1	-0.007 / 0.00	1.00
12L/d768 classifier (s0)	sex	179	87	-0.000 [-0.001, 0.001]	0.000 [0.000, 0.042]	0	-0.015 / **0.01**	1.00
12L/d768 classifier (s2)	sex	179	91	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	race	59	33	+0.001 [-0.006, 0.011]	0.000 [0.000, 0.104]	0	-0.015 / 0.00	1.00
12L/d768 classifier (s2)	race	59	34	-0.001 [-0.001, -0.000]	0.000 [0.000, 0.102]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	89	+0.001 [-0.001, 0.002]	0.000 [0.000, 0.041]	0	-0.015 / **0.01**	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	101	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.037]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	119	+0.000 [-0.001, 0.001]	0.000 [0.000, 0.031]	0	-0.015 / **0.01**	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	119	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.031]	1	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	+0.001 [-0.000, 0.002]	0.000 [0.000, 0.036]	0	-0.015 / **0.01**	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	104	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.036]	0	-0.000 / 0.00	1.00
12L/d768 classifier (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	+0.000 [-0.001, 0.002]	0.000 [0.000, 0.042]	0	-0.015 / **0.01**	1.00
12L/d768 classifier (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	91	-0.000 [-0.000, 0.000]	0.000 [0.000, 0.041]	0	-0.000 / 0.00	1.00
12L/d768 generative (s0)	sex	179	91	-0.000 [-0.001, 0.000]	0.000 [0.000, 0.041]	0	+1.366 / **1.00**	1.00
12L/d768 generative (s2)	sex	179	84	+0.000 [-0.000, 0.000]	0.000 [0.000, 0.044]	0	-0.001 / 0.00	1.00
12L/d768 generative (s0)	race	59	33	-0.000 [-0.000, -0.000]	0.000 [0.000, 0.104]	0	+1.340 / **1.00**	1.00
12L/d768 generative (s2)	race	59	31	+0.000 [0.000, 0.001]	0.000 [0.000, 0.110]	6	-0.000 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	98	+0.001 [-0.004, 0.006]	0.000 [0.000, 0.038]	2	+1.378 / **1.00**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \leq 5$	182	88	+0.003 [0.000, 0.005]	0.000 [0.000, 0.042]	0	-0.001 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	117	+0.001 [-0.003, 0.005]	0.000 [0.000, 0.032]	0	+1.375 / **1.00**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 5$	219	110	-0.001 [-0.002, 0.000]	0.000 [0.000, 0.034]	1	-0.001 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	103	-0.001 [-0.006, 0.003]	0.000 [0.000, 0.036]	1	+1.397 / **1.00**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 10$	199	96	+0.001 [-0.003, 0.005]	0.000 [0.000, 0.038]	0	-0.001 / 0.00	1.00
12L/d768 generative (s0)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	90	+0.002 [-0.001, 0.006]	0.000 [0.000, 0.041]	0	+1.412 / **1.00**	1.00
12L/d768 generative (s2)	$\lvert\Delta\text{YOB}\rvert \geq 15$	175	88	+0.001 [-0.003, 0.006]	0.000 [0.000, 0.042]	0	-0.001 / 0.00	1.00

D.3 COGNITIVE-SCORE ACTIVATION SWAPPING

For numerical cognitive and functional variables, we identify the token span containing the feature value and transplant the donor's residual-stream state at that position into an opposite-label recipient. Swaps are performed layer by layer in both directions. The primary cohort matches participants on sex and year of birth; analyses are available for both a ± 5-year caliper and an exact-YOB cohort. We report the change in $d(x)$ and the prediction-flip rate. A final-residual read-out patch serves as a liveness check.

This experiment deliberately changes an *internal representation* of a number rather than the number in the text. It therefore complements the input-level counterfactual experiments below.

D.3.1 COGNITIVE-SCORE ACTIVATION SWAPPING RESULTS

For most numerical cognitive and functional features, swapping the feature- position activation between matched opposite-label participants changes the class margin without reversing the predicted label. The notable exception is the 12-layer, $d = 768$ generative control: swapping the layer-0 representation of FAQ Total produces approximately 60% CN$\rightarrow$AD flips and approximately 17% AD$\rightarrow$CN flips in the reported cohort. The second most influential feature is 'FAQ Taxes & Papers', which achieves a 12% flip rate CN$\rightarrow$AD, but the effect is not symmetrical. This may be an early sensitive sign of cognitive decline. Across features, the largest margin effects tend to occur at Layer 0 or within the first few layers and then decay at the original feature position, consistent with information being propagated away from that location rather than with the feature becoming irrelevant downstream.

Table 18: Cognitive-feature activation swapping on the from-scratch toys: the donor's activations at one feature's value token are patched into the recipient. *max flip* is the largest label-flip rate over all 25 features and layers *within that direction*, measured over n *gated* – the pairs the model already classifies correctly on both sides, the only ones where a flip is possible, which is why *ceiling* is 1.00. *peak* Δ is over all 403 pairs instead: the gate is defined by the sign of the baseline margin and the change is measured against that same baseline, so gating inflates it by a factor tracking the cohort shrinkage – selection, not signal. *ceiling* is the flip rate when the whole residual stream is overwritten at the read-out token, which bounds any intervention there and is feature-independent. *active* counts features whose flip rate exceeds 0.05 anywhere. The gap between max flip and ceiling is the result: patching a single numeric value token moves almost nothing, while the same position overwritten wholesale moves most of the cohort. The two directions are reported separately because they are not symmetric – CN→AD is the larger of the two in 11 of the 12 runs – and because they move the margin in opposite directions by construction, so a single signed peak pooled over both would not be comparable down the column. Cohort: 403 exact year-of-birth and sex matched pairs.

Model	Seed	n gated	best feature	peak Δ	max flip	layer	active	ceiling
AD→CN — normalising edits, patients called AD								
6L d256 clf	0	261	AVLT_Trial5_Total	-0.09	0.02	0	0/25	1.00
6L d256 clf	2	258	FAQ_TaxesPapers	-0.02	0.00	0	0/25	1.00
6L d256 gen	0	254	FAQ_TaxesPapers	-0.01	0.00	0	0/25	1.00
6L d256 gen	2	272	FAQ_Total	-0.53	0.00	0	0/25	1.00
12L d256 clf	0	264	TrailMaking_PartB_Time	-0.00	0.00	1	0/25	1.00
12L d256 clf	2	269	FAQ_TaxesPapers	-0.02	0.00	0	0/25	1.00
12L d256 gen	0	272	AVLT_Trial5_Total	-0.10	0.01	0	0/25	1.00
12L d256 gen	2	266	FAQ_TaxesPapers	-0.02	0.02	0	0/25	1.00
12L d768 clf	0	264	FAQ_TaxesPapers	-0.05	0.01	0	0/25	1.00
12L d768 clf	2	270	TrailMaking_PartB_Time	-0.00	0.00	0	0/25	1.00
12L d768 gen	0	272	FAQ_Total	-1.15	**0.17**	0	1/25	1.00
12L d768 gen	2	261	FAQ_Dates	-0.01	0.01	0	0/25	1.00
CN→AD — impairing edits, patients called CN								
6L d256 clf	0	261	FAQ_Dates	+0.10	0.03	0	0/25	1.00
6L d256 clf	2	258	FAQ_ChecksBills	+0.14	0.03	0	0/25	1.00
6L d256 gen	0	254	FAQ_TaxesPapers	+0.16	0.02	0	0/25	1.00
6L d256 gen	2	272	FAQ_Total	+0.12	0.08	0	1/25	1.00
12L d256 clf	0	264	TrailMaking_PartA_Time	+0.02	0.00	1	0/25	1.00
12L d256 clf	2	269	FAQ_TaxesPapers	+0.12	0.01	0	0/25	1.00
12L d256 gen	0	272	AVLT_Trial5_Total	+0.13	0.03	0	0/25	1.00
12L d256 gen	2	266	FAQ_ChecksBills	+0.11	0.08	0	1/25	1.00
12L d768 clf	0	264	FAQ_TaxesPapers	+0.26	0.12	0	7/25	1.00
12L d768 clf	2	270	TrailMaking_PartB_Time	+0.00	0.00	0	0/25	1.00
12L d768 gen	0	272	FAQ_Total	+0.98	**0.99**	0	1/25	1.00
12L d768 gen	2	261	FAQ_TaxesPapers	+0.13	0.01	0	0/25	1.00

D.4 Input-level counterfactual perturbations

Each FAQ item in the narratives encodes severity twice, as a verbal descriptor and as a parenthesised digit (e.g. `Normal (0)`), and the FAQ total is reported as a separate digit. Because the two encodings are redundant, they can be edited independently. We use targeted text rewrites to test which encoding the models rely on. Every rewrite moves a narrative toward the opposite class: AD narratives receive normalising edits (AD→CN), and CN narratives receive impairing edits (CN→AD).

Numeric channel. These edits leave every descriptor untouched and set digits to the median of the opposite class, at three scopes: (i) the FAQ total alone; (ii) the scores of the ten FAQ items, excluding the total; and (iii) the totals of all sixteen test measures, i.e. the FAQ total plus fifteen other cognitive scores (listed in Table 19). Scopes (i) and (iii) share the FAQ total, so if (iii) flips no more participants than (i), the model's numeric sensitivity is carried by the FAQ total alone. Scope (ii) tests whether item-level scores are read independently of their sum.

Clinical severity channel. The descriptors of all ten FAQ items are rewritten to the extreme of the opposite class (`Normal` for AD→CN, `Dependent` for CN→AD) with all digits unchanged, so that words and numbers disagree. A second edit rewrites descriptors and digits together (`Normal (0)` with FAQ total 0, or `Dependent (3)` with FAQ total 30), giving a consistent reference.

Content versus length. Descriptor edits change length as well as meaning. To separate the two, a score-preserving paraphrase replaces `Normal (0)` with the equivalent-score but longer `Never did, but could do now (0)`, adding 56 tokens while leaving every score unchanged. A length-matched control instead inserts 56 tokens of clinically neutral filler. Conversely, `Normal (0)` and `Dependent (3)` differ strongly in severity but only modestly in length. Together these edits separate wording, severity, and length.

Negative control. The semantic edit is applied in the direction of the participant's current label (e.g. a CN narrative is made maximally normal on the FAQ). It should not flip any prediction, and it sets the floor against which the other flip rates are read.

Cohort and outcome. All edits are evaluated on the exact year-of-birth and sex matched AD–CN pairs in which both members are correctly classified. The AD members form the AD→CN set and the CN members the CN→AD set, so n is fixed per model across directions and conditions ($n = 266$–317). The outcome is the flip rate, the proportion of participants whose predicted label changes relative to their intact narrative, reported with Wilson 95% intervals.

D.4.1 INPUT-COUNTERFACTUAL RESULTS

Table 19 gives flip rates for 16 runs in both directions. The findings are summarised below by channel. For each channel we state whether the effect generalises across architecture, training history (from scratch vs. pretrained) and objective (classifier vs. generative), or depends on the seed.

The verbal channel dominates across all models. In the CN→AD direction, rewriting the FAQ descriptors alone flips at least 0.84 of participants in 14 of 16 runs, and 0.98–1.00 in 12. When descriptors and digits are rewritten together, the flip rate is at least 0.88 in all 16 runs. This reliance on wording generalises across architectures, both training histories and both objectives, with the exception of 12L/d768 generative seed 0 and 1.4B classifier seed 0, which have more numerical sensitivity. The normalising direction is weaker and more variable: descriptor rewrites flip 0.41–1.00 of AD participants in 13 of 16 runs. Objective has no consistent effect here. The 6L/d256 generative runs flip more than the classifiers (0.75, 0.89 vs. 0.49, 0.54), but the ordering reverses for 12L/d256 (classifiers 0.99–1.00 vs. generative 0.45–0.49). The asymmetry between directions likely reflects the conditional nature of the interventions: each edit is evaluated conditional on the rest of the clinical summary, and this residual context differs systematically between cohorts. When the FAQ section of an AD narrative is normalised, the summary retains impaired cognitive test scores, so the model still has positive evidence for impairment; the edit removes one source of signal rather than the signal itself. Adding severe functional impairment to a CN narrative, by contrast, introduces positive evidence of impairment against a residual context that is merely unremarkable, and functional decline with relatively preserved test performance remains a plausible impaired presentation. This mirrors clinical practice, where a few impaired features raise suspicion of AD in a CN individual, but revising an established AD diagnosis requires the full picture to normalise. Under the residual-evidence explanation, AD→CN flip rates should approach the CN→to-AD rates once no contradicting evidence remains; a persistent gap under full normalisation would instead implicate model-level asymmetries, such as a default bias toward the impaired label or an asymmetric decision threshold. The all20ceiling and FAQall experiments in Table 20 (Appendix D.5.1) change all verbal descriptors achieving full normalisation/impairement outside of numerical scores. This full verbal joint normalisation restores AD→CN flip rates to near the CN→AD level for most runs (0.86–1.00 vs 1.00), supporting the residual-evidence account for the typical asymmetry; however, in a minority of runs (1.4B classifier seed 0, 12L d768 generative seed 0, and 1.4B generative seed 2), near-zero flip rates persist, which confirms they are more sensitive to numerical features.

Reliance on digits is rare and seed-specific. Rewriting the FAQ total digit alone flips at most 0.10 of participants in 28 of 32 models. Three runs read it: the 12L/d768 generative model at seed

0 (1.00 CN→AD, 0.17 AD→CN), the Pythia-1.4B classifier at seed 0 (0.95, 0.10) and the Pythia-1.4B generative model at seed 2 (0.96, 0.04). The partner seed of each is numerically inert ($\leq$0.02). The effect therefore reflects the individual training run rather than architecture, objective or training history, which it spans. The only regularity is that it appears in the two largest models and never in 6L/d256 or 12L/d256. The same three runs are the least responsive to descriptor rewrites (0.01–0.18 AD→CN; 0.26, 0.33 and 0.84 CN→AD), so reliance on numbers comes at the expense of reliance on words. In the 12L/d768 generative (seed 0) and Pythia-1.4B generative (seed 2) runs, rewriting all sixteen totals adds almost nothing over the FAQ total (0.18 vs. 0.17 and 1.00 vs. 1.00; 0.06 vs. 0.04 and 1.00 vs. 0.96), so their numeric sensitivity is carried by the FAQ total. The 12L/d768 result agrees with the independent activation-swapping analysis. The Pythia-1.4B classifier (seed 0) is the exception: in the AD→CN direction the 16-total edit flips 0.34 of participants against 0.10 for the FAQ total, so other test totals contribute there. None of the three runs responds to the ten FAQ item digits ($\leq$0.05); they read the summary score, not its components. A disjoint set of three runs shows the reverse pattern, responding to the item digits but not the total, and only in the CN→AD direction. These are the 6L/d256 classifier at seed 2 (0.58), the 12L/d256 generative model at seed 2 (0.73) and the 12L/d768 classifier at seed 0 (0.19). Their FAQ total sensitivity is $\leq$0.06.

Score-preserving rewording flips diagnoses, but length does not. Replacing `Normal (0)` with `Never did, but could do now (0)` leaves every score unchanged, yet flips CN participants to AD in 14 of 16 runs (0.25–1.00), and in 8 runs flips $\geq$0.89 of them. A clinically neutral 56-token filler that matches the length increase flips $\leq$ 0.01 in 15 of 16 runs. Therefore, wording, not length, drives the effect. The effect is present in every architecture and is largest in seed-2 runs: seed 2 flips at least 0.99 of participants for six of the eight architecture–objective combinations, whereas the seed-0 counterparts range from 0.01 to 0.92. The exceptions are the two 12L/d768 models (0.38 and 0.63 at seed 2), which are also the least sensitive overall (at most 0.63). Pretraining-and-LoRA-adaptation confers no protection: all four Pythia-1.4B runs flip CN→AD 0.73–1.00 in the clinically invariant setting with surface rephrasing. We hypothesise that this sensitivity may be due to lexical similarity between the clinically equivalent `Never did, but could do now (0)` and the impaired descriptor `Never did, would have difficulty now (1)`, leading the models to interpret the shared phrasing as impairment rather than the score it carries. Further experiments could characterise if this sensitivity is driven by lexical, syntactic, or distributional cues.

Clinically empty text is inert, except in one pretrained run. Length-matched filler flips at most 0.02 of participants in 30 of 32 runs, and the negative control flips no participant in 30 of 32 rows. This indicates that, in general, the models respond to clinical content rather than to arbitrary changes in the text. The single outlier is the Pythia-1.4B generative model at seed 0. Filler flips 0.33 of its AD participants and 0.59 of its CN participants, the negative control flips 0.075 [0.05, 0.11], and score-preserving rewording flips 0.07 of AD participants. It is the only run whose predictions move under edits that carry no diagnostic information. Despite its high in-distribution accuracy, its input sensitivity therefore rests partly on clinically unsound features. The only other negative-control flip is a single participant (1 of 313) for the 12L/d768 classifier at seed 0.

Table 19: Flip rates under counterfactual text rewrites, in both directions (seeds s0 and s2). Each narrative is rewritten and the model re-run, and a flip is counted when the predicted label changes. Cells give the flip rate with a Wilson 95% interval. Cohort: exact year-of-birth and sex matched AD–CN pairs with both members correctly classified, so n is fixed per model across directions and conditions. The two halves use mirror-image edits, normalising for AD→CN and impairing for CN→AD. The *numeric channel* rewrites digits only, leaving every descriptor untouched, at three scopes, each set to the opposite-class median: the FAQ total; the score on each of the ten FAQ items; and the totals of sixteen test measures (FAQ, both clock scores, nine AVLT measures, category fluency, both Trail Making times, ANART errors). The *clinical channel* rewrites the FAQ clinical severity descriptors, either alone or together with all FAQ digits. *Content vs. length* rewords descriptors at constant clinical severity score, or inserts clinically empty filler of the same token count. The negative control (the semantic edit applied toward the current label) is omitted for width. It flips no participant in 30 of 32 rows; the exceptions are Pythia-1.4B generative at s0 (0.075 [0.05, 0.11]) and the 12L/d768 classifier at s0 (one participant of 313).

AD→CN — patients the model calls AD; can a normalising edit flip them to CN

		numeric channel			clinical channel		content vs. length	
Condition	n	FAQ Total digit	10 FAQ digits	16 test totals	FAQ words →Normal	Normal (0), total 0	same score, +56 tok	filler +56 tok
6L/d256 classifier (s0)	303	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]	0.06 [0.04, 0.09]	0.49 [0.43, 0.54]	0.66 [0.61, 0.71]	0.00 [0.00, 0.01]	0.01 [0.00, 0.03]
6L/d256 classifier (s2)	298	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]	0.01 [0.00, 0.02]	0.54 [0.48, 0.59]	0.82 [0.77, 0.86]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
6L/d256 generative (s0)	295	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.01 [0.00, 0.02]	0.75 [0.69, 0.79]	0.90 [0.86, 0.93]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
6L/d256 generative (s2)	313	0.01 [0.00, 0.02]	0.00 [0.00, 0.01]	0.01 [0.00, 0.03]	0.89 [0.85, 0.92]	0.97 [0.95, 0.98]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d256 classifier (s0)	308	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d256 classifier (s2)	313	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]	0.01 [0.00, 0.02]	0.99 [0.97, 1.00]	0.99 [0.98, 1.00]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d256 generative (s0)	310	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.02 [0.01, 0.04]	0.49 [0.43, 0.54]	0.86 [0.82, 0.90]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d256 generative (s2)	310	0.00 [0.00, 0.02]	0.00 [0.00, 0.02]	0.04 [0.02, 0.07]	0.45 [0.39, 0.50]	0.76 [0.71, 0.80]	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]
12L/d768 classifier (s0)	313	0.00 [0.00, 0.01]	0.02 [0.01, 0.04]	0.03 [0.01, 0.05]	0.41 [0.35, 0.46]	0.75 [0.70, 0.79]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d768 classifier (s2)	315	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.98 [0.96, 0.99]	0.98 [0.96, 0.99]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
12L/d768 generative (s0)	317	0.17 [0.14, 0.22]	0.00 [0.00, 0.01]	0.18 [0.14, 0.22]	0.01 [0.00, 0.03]	0.97 [0.94, 0.98]	0.01 [0.00, 0.03]	0.00 [0.00, 0.01]
12L/d768 generative (s2)	308	0.00 [0.00, 0.01]	0.02 [0.01, 0.04]	0.00 [0.00, 0.02]	0.98 [0.96, 0.99]	1.00 [0.99, 1.00]	0.00 [0.00, 0.02]	0.00 [0.00, 0.02]
1.4B classifier (s0)	301	0.10 [0.07, 0.14]	0.00 [0.00, 0.01]	0.34 [0.28, 0.39]	0.06 [0.04, 0.09]	0.30 [0.25, 0.36]	0.00 [0.00, 0.01]	0.02 [0.01, 0.04]
1.4B classifier (s2)	304	0.01 [0.01, 0.03]	0.00 [0.00, 0.02]	0.01 [0.01, 0.03]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
1.4B generative (s0)	266	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]	0.00 [0.00, 0.02]	0.99 [0.97, 1.00]	1.00 [0.98, 1.00]	0.07 [0.05, 0.11]	0.33 [0.28, 0.39]
1.4B generative (s2)	301	0.04 [0.03, 0.07]	0.00 [0.00, 0.01]	0.06 [0.04, 0.10]	0.18 [0.14, 0.23]	0.99 [0.98, 1.00]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]

CN→AD — patients the model calls CN; can an impairing edit flip them to AD

		numeric channel			clinical channel		content vs. length	
Condition	n	FAQ Total digit	10 FAQ digits	16 test totals	FAQ words →Dependent	Dependent (3), total 30	same score, +56 tok	filler +56 tok
6L/d256 classifier (s0)	303	0.03 [0.02, 0.06]	0.00 [0.00, 0.01]	0.11 [0.08, 0.15]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.92 [0.88, 0.94]	0.00 [0.00, 0.01]
6L/d256 classifier (s2)	298	0.01 [0.00, 0.03]	0.58 [0.52, 0.64]	0.12 [0.09, 0.16]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.00 [0.00, 0.02]
6L/d256 generative (s0)	295	0.00 [0.00, 0.01]	0.06 [0.04, 0.09]	0.02 [0.01, 0.04]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.68 [0.63, 0.73]	0.00 [0.00, 0.01]
6L/d256 generative (s2)	313	0.08 [0.05, 0.11]	0.08 [0.05, 0.11]	0.08 [0.05, 0.11]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.00 [0.00, 0.02]
12L/d256 classifier (s0)	308	0.00 [0.00, 0.01]	0.00 [0.00, 0.02]	0.00 [0.00, 0.01]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.01 [0.00, 0.02]	0.00 [0.00, 0.01]
12L/d256 classifier (s2)	313	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.00 [0.00, 0.01]
12L/d256 generative (s0)	310	0.00 [0.00, 0.02]	0.08 [0.05, 0.11]	0.03 [0.02, 0.05]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.25 [0.21, 0.31]	0.00 [0.00, 0.01]
12L/d256 generative (s2)	310	0.06 [0.04, 0.09]	0.73 [0.68, 0.78]	0.21 [0.17, 0.26]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.00 [0.00, 0.02]
12L/d768 classifier (s0)	313	0.00 [0.00, 0.01]	0.19 [0.15, 0.24]	0.06 [0.04, 0.10]	0.98 [0.95, 0.99]	1.00 [0.99, 1.00]	0.40 [0.34, 0.45]	0.01 [0.00, 0.03]
12L/d768 classifier (s2)	315	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.38 [0.33, 0.44]	0.00 [0.00, 0.01]
12L/d768 generative (s0)	317	1.00 [0.99, 1.00]	0.01 [0.00, 0.03]	1.00 [0.99, 1.00]	0.26 [0.22, 0.31]	1.00 [0.99, 1.00]	0.02 [0.01, 0.04]	0.01 [0.00, 0.02]
12L/d768 generative (s2)	308	0.00 [0.00, 0.01]	0.04 [0.02, 0.07]	0.00 [0.00, 0.01]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.63 [0.57, 0.68]	0.00 [0.00, 0.01]
1.4B classifier (s0)	301	0.95 [0.92, 0.97]	0.05 [0.03, 0.08]	1.00 [0.99, 1.00]	0.33 [0.28, 0.38]	1.00 [0.99, 1.00]	0.89 [0.85, 0.92]	0.00 [0.00, 0.02]
1.4B classifier (s2)	304	0.02 [0.01, 0.04]	0.02 [0.01, 0.05]	0.03 [0.02, 0.06]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	1.00 [0.98, 1.00]	0.01 [0.01, 0.03]
1.4B generative (s0)	266	0.00 [0.00, 0.02]	0.05 [0.03, 0.08]	0.00 [0.00, 0.02]	0.88 [0.84, 0.91]	0.88 [0.84, 0.92]	0.73 [0.68, 0.78]	0.59 [0.53, 0.65]
1.4B generative (s2)	301	0.96 [0.94, 0.98]	0.01 [0.00, 0.03]	1.00 [0.98, 1.00]	0.84 [0.79, 0.87]	1.00 [0.98, 1.00]	0.99 [0.97, 1.00]	0.00 [0.00, 0.02]

D.5 SINGLE-ITEM SUFFICIENCY AND LEAVE-ONE-OUT NECESSITY

Since the verbal input counterfactuals produced overall the strongest effects, we investigate which verbal features are necessary and sufficient to flip the diagnosis. The verbal intervention includes 20 items: ten FAQ items, five Clock Drawing items, and five Clock Copying items. The Clock Copying Items are used as negative control as we know these are not represented by the models in the linear probing experiment.

For each direction, only participants that the model classifies correctly are included. The principal cohort uses exact year-of-birth and sex matched opposite-label pairs (n=403). For item j, single-item

sufficiency is the flip rate when only that item is rewritten toward the opposite class,

$$\mathrm{Suff}_j = P\Big(\hat{y}(x^{(j)}) \neq \hat{y}(x) \mid x^{(j)} \neq x\Big), \tag{10}$$

where conditioning on an actual edit prevents a pre-existing normal value from being counted as evidence that the feature is ineffective.

Let F_{all} be the flip rate after rewriting all 20 verbal items and F_{-j} the rate when all items except j are rewritten. Per-item necessity is

$$\mathrm{Nec}_j = F_{\mathrm{all}} - F_{-j}. \tag{11}$$

Necessity is not interpreted for a model/direction whose all-item ceiling is too low to support a meaningful drop. Negative-control rewrites alter handedness, marital status, and retirement status. FAQ-specific score-preserving length controls replace `Normal (0)` with `Never did, but could do now (0)`.

D.5.1 Single-item sufficiency and leave-one-out results

Single-feature interventions show a pronounced directional asymmetry. In the CN→AD direction, several FAQ items are individually highly sufficient in the 6-layer controls and in Pythia-1.4B generative. In the supplied table, assembling tax records, heating water, preparing a balanced meal, remembering appointments, and writing checks each reach flip rate 1.00 in the 6-layer generative control and in Pythia-1.4B generative. Clock Drawing items are substantially weaker and Clock Copying items are uniformly ineffective in these single-item interventions, as expected. Conversely, no individual normalising edit is strongly sufficient for AD→CN, consistent with broader redundancy of abnormal findings once an AD prediction has been established.

Single-item interventions reveal a pronounced directional asymmetry, and a spectrum of integration styles. In the CN→AD direction, the narrow from-scratch models are switch-like: both 12L/d=256 classifier seeds and the 12L/d=768 classifier at seed 2 flip 1.00 of matched CN patients on every single FAQ item, and the 6L generative model saturates on most (7–8 of 10 items at ≥ 0.99 per seed). The 6L and 12L/d768 classifiers at seed 0 are graded instead (0.17–0.77), and the pretrained 1.4B models are integrators: no single item exceeds 0.44 in either objective at either seed, with one idiosyncratic exception, travelling out at 0.94 in the classifier at seed 2, despite all-item ceilings of 0.55–1.00.

The most sufficient FAQ items are assembling tax records (0.44–1.00), shopping alone (0.66–1.00), preparing a balanced meal (0.31–0.96) and remembering appointments (0.34–0.91). Therefore, the inability to perform these is a sensitive markers to models of an AD diagnosis.

Clock items are not as significant: clock-drawing (CDT) items reach at most 0.19 in sufficiency for 12L/d768 model, and clock-copying (CCT) items are 0.00 for every model. The copying task carries no causally used diagnostic signal anywhere in this set, as expected, as this feature is not represented in our linear probing experiment.

The necessity block shows that within the FAQ items, assembling tax records ($F_{\mathrm{all}} - F_{-j} = 0.55$–$0.72$), remembering appointments (0.52–0.69) and writing checks (0.32–0.49) are the most necessary overall ($F_{\mathrm{all}} - F_{-j} = 0.52$–$0.72$) and heating water the least (0.06–0.13), across both objectives, all architectures, and both seeds. Given the seed-dependence of circuits and channels elsewhere in this work, the agreement of eleven independently trained solutions on which clinical items anchor the AD-CN prediction is the strongest invariant in the causal analyses.

The verbally-blind models are internally consistent: the 12L/d=768 generative model at seed 0 is unmoved by any verbal edit in either direction (sufficiency ≤ 0.07; necessity ceiling 0.01), matching its numeric-channel profile, and the 1.4B classifier at seed 0 supports no necessity estimate (ceiling 0.06) despite partial sufficiency, an asymmetry between directions rather than uniform blindness.

Table 20: Per-item sufficiency and necessity. *Sufficiency* is the flip rate when that item alone is rewritten, conditioned on the rewrite changing the text; *necessity* is $F_{\text{all}} - F_{-j}$, the all20 ceiling minus the flip rate that survives when every item except j is rewritten. The two are reported in opposite directions because each is degenerate in the other, so each block carries its own n and ceiling. Clock items are prefixed CDT (drawing) or CCT (copying); the rest are FAQ. Every item is verbal. The italic *FAQ all* and *Clock all* rows treat the ten FAQ items and the ten clock items (CDT+CCT) as single units. Cohort: exact age+sex matched pairs, n pairs per model

Sufficiency — CN→AD, item rewritten alone

Condition	n	all20 ceiling	FAQ all	Clock all	Assembling tax	Heating water	Keeping track current affairs	Paying attention TV	Playing a game	Preparing balanced meal	Remembering appointments	Shopping alone	Traveling out	Writing checks	CDT circular_face	CDT hands_time	CDT numbers	CDT symmetry	CDT two_hands	CCT circular_face	CCT hands_time	CCT numbers	CCT symmetry	CCT two_hands
6L d256 s0 classifier	303	1.00	1.00	0.43	0.54	0.32	0.32	0.40	0.48	0.73	0.77	0.66	0.17	0.51	0.06	0.05	0.05	0.05	0.04	0.00	0.00	0.00	0.00	0.00
6L d256 s2 classifier	298	1.00	1.00	0.05	0.95	0.49	0.49	0.34	0.70	0.96	0.60	1.00	0.17	0.47	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
12L d256 s0 classifier	308	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s2 classifier	313	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d768 s0 classifier	313	1.00	1.00	0.19	0.71	0.39	0.33	0.34	0.36	0.31	0.34	0.69	0.34	0.72	0.16	0.19	0.18	0.17	0.16	0.00	0.01	0.00	0.00	0.00
12L d768 s2 classifier	315	1.00	1.00	0.09	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.09	0.11	0.10	0.10	0.09	0.00	0.00	0.00	0.00	0.00
1.4B classifier	301	0.55	0.55	0.02	0.06	0.04	0.05	0.04	0.03	0.06	0.06	0.04	0.10	0.06	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
1.4B classifier	304	1.00	1.00	0.02	0.13	0.10	0.11	0.12	0.09	0.10	0.21	0.12	0.94	0.09	0.02	0.02	0.01	0.02	0.02	0.00	0.01	0.00	0.00	0.00
6L d256 s0 generative	295	1.00	1.00	0.10	1.00	1.00	0.91	0.77	1.00	1.00	1.00	1.00	0.59	1.00	0.01	0.01	0.01	0.02	0.02	0.00	0.00	0.00	0.00	0.00
6L d256 s2 generative	313	1.00	1.00	0.01	1.00	1.00	1.00	0.21	1.00	1.00	0.99	1.00	1.00	0.15	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s0 generative	310	1.00	1.00	0.04	0.44	0.48	0.45	0.58	0.89	0.43	0.91	0.57	0.38	0.77	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s2 generative	310	1.00	1.00	0.40	0.06	0.11	0.09	0.98	0.75	0.12	0.51	0.06	0.06	0.30	0.06	0.12	0.07	0.07	0.12	0.00	0.00	0.00	0.00	0.00
12L d768 s0 generative	317	1.00	0.72	0.16	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.07	0.01	0.01	0.03	0.00	0.00	0.00	0.00	0.00
12L d768 s2 generative	308	1.00	1.00	0.00	0.03	0.05	0.09	0.10	0.08	0.10	0.10	0.05	0.05	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.4B generative	266	0.92	0.88	0.18	0.02	0.09	0.09	0.09	0.08	0.05	0.06	0.06	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.00	0.00
1.4B generative	301	0.93	0.90	0.05	0.16	0.23	0.33	0.44	0.19	0.36	0.37	0.16	0.14	0.10	0.02	0.03	0.02	0.02	0.03	0.00	0.00	0.00	0.00	0.00

Necessity — AD→CN, all items EXCEPT this one rewritten

Condition	n	all20 ceiling	FAQ all	Clock all	Assembling tax	Heating water	Keeping track current affairs	Paying attention TV	Playing a game	Preparing balanced meal	Remembering appointments	Shopping alone	Traveling out	Writing checks	CDT circular_face	CDT hands_time	CDT numbers	CDT symmetry	CDT two_hands	CCT circular_face	CCT hands_time	CCT numbers	CCT symmetry	CCT two_hands
6L d256 s0 classifier	303	0.86	0.83	0.20	0.52	0.08	0.27	0.30	0.28	0.29	0.55	0.31	0.30	0.40	0.00	0.08	0.01	0.05	0.01	0.00	0.00	0.00	0.00	0.00
6L d256 s2 classifier	298	0.87	0.87	0.05	0.57	0.08	0.28	0.32	0.32	0.29	0.59	0.31	0.32	0.40	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s0 classifier	308	1.00	1.00	0.00	0.69	0.13	0.37	0.42	0.40	0.36	0.72	0.40	0.49	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s2 classifier	313	0.99	0.99	0.00	0.69	0.13	0.36	0.41	0.39	0.36	0.70	0.40	0.49	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12L d768 s0 classifier	313	0.98	0.94	0.23	0.60	0.12	0.33	0.36	0.38	0.32	0.62	0.38	0.37	0.52	0.01	0.17	0.05	0.12	0.03	0.00	0.00	0.00	0.00	0.00
12L d768 s2 classifier	315	1.00	0.96	0.02	0.66	0.12	0.35	0.39	0.37	0.34	0.68	0.38	0.47	0.54	0.00	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
1.4B classifier	301	0.06	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1.4B classifier	304	1.00	1.00	0.00	0.56	0.06	0.27	0.35	0.26	0.28	0.66	0.32	0.46	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6L d256 s0 generative	295	0.97	0.96	0.07	0.66	0.12	0.33	0.37	0.37	0.33	0.68	0.37	0.43	0.53	0.00	0.05	0.02	0.03	0.01	0.00	0.00	0.00	0.00	0.00
6L d256 s2 generative	313	0.94	0.94	0.01	0.63	0.12	0.34	0.38	0.26	0.33	0.66	0.38	0.45	0.35	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00
12L d256 s0 generative	310	0.95	0.95	0.09	0.63	0.12	0.34	0.38	0.37	0.33	0.68	0.38	0.43	0.52	0.00	0.07	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00
12L d256 s2 generative	310	0.91	0.89	0.16	0.19	0.06	0.09	0.36	0.38	0.14	0.60	0.08	0.08	0.39	-0.00	0.12	0.02	0.03	0.02	0.00	0.00	0.00	0.00	0.00
12L d768 s0 generative	317	0.01	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
12L d768 s2 generative	308	1.00	1.00	0.00	0.32	0.11	0.35	0.43	0.29	0.21	0.66	0.34	0.37	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.4B generative	266	1.00	0.95	0.00	0.01	0.01	0.01	0.03	0.02	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.4B generative	301	0.23	0.22	0.04	0.09	0.01	0.02	0.02	0.06	0.06	0.08	0.03	0.03	0.02	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00

D.6 GREEDY MINIMAL SUFFICIENT SETS AND SET-LEVEL NECESSITY

A true greedy search is used rather than accumulating items according to their single-item ranking. At each step, every remaining candidate is rescored in the context of the already selected subset and the candidate that maximises flip rate on a fixed search subset is added. The selected set is then evaluated on the full cohort. The search stops only when the *full-cohort* flip rate reaches 0.90 or all 20 items have been selected.

For a resulting set S, set-level sufficiency is its full-cohort flip rate $F(S)$. Element-wise set necessity is

$$\Delta_s = F(S) - F(S \setminus \{s\}), \qquad s \in S. \tag{12}$$

If the greedy procedure never reaches the sufficiency threshold, necessity within that set is reported as undefined rather than interpreted.

Two controls distinguish feature identity from subset cardinality. First, we draw size-matched random subsets R from the same 20-item universe and report the one-sided empirical probability that $F(R) \geq F(S)$. Second, for each $s \in S$ we replace s with the best item outside S, where the substitute is selected on the search subset and evaluated on the full cohort. Recovery of $F(S)$ after substitution indicates that an item of that *type* may be needed without the original feature being uniquely special.

D.6.1 GREEDY MINIMAL-SET RESULTS

The discrepancy between sufficiency and necessity has motivated the greedy search of the minimum necessary and sufficient set of features jointly edited together to reach >90% flip rate (or until all 20 features run out).

The current greedy sequences reinforce the directional asymmetry of the single-item analysis. For CN→AD, every reported model reaches a flip rate of at least 0.90 after one to eight selected edits. The only exception is seed 0 1.4B classifier, which achieves only 55% flip rate when utilising the whole 20 feature set. The 6-layer generative and 12L/256 classifier control reach 1.00 with a single selected feature in the reported cohort, while the 6-layer classifier conditions 1.00 after two edits and the wide 12-layer classifier and generative control requires between 1 and 5 features, and 1.4B requires increasingly 1 to 20 features. Therefore overall the number of impairing features required increases with model size, but it is not obviously affected by task formulation.

In the opposite AD→CN direction, no reported model reaches 0.90 within seven greedy edits; final flip rates after six steps range from 0.06 to 1.00. This is consistent with a decision rule in which a small amount of introduced impairment can be sufficient for an AD prediction, whereas normalising an established AD prediction requires a broader change to the clinical profile.

The 6-layer generative controls illustrates the directional asymmetry: one selected impairment-inducing item achieves a 100% CN-to-AD flip rate, whereas 10–11 normalising edits are needed to reach 92% in the opposite direction. Random subsets do not recover the effect of the greedy subset in the normalising direction (AD to CN), which eliminates the possibility of the effect being simply due to cardinality. Moreover, substituting a single item from the greedy subset is statistically significantly worse for the majority of models, suggesting a specific sequence of normalising features (Table 22). In contrast, for the impairment-inducing changes (CN→AD), the greedy subset is overall small for the majority of the from-scratch control models, and substitution of a single item within the greedy subset does not significantly drop their flip rate effect, suggesting that the greedy minimal subset is not unique for these models. Significant drop in performance with substitution, which implies uniqueness of the impairing faetures, is seen only in three classifier models, namely 6L (Shopping) and 12L narrow (Checks) scratch control models and 1.4B (Travel) model, with significant seed dependence (Table 21)

Table 21: Greedy minimal sufficient subset. Each step rescores all 20 items for that model and adds the best, so the subset is model-specific in identity and order – step 1 is not a shared feature. *final* is the flip rate at the last step; short rows are early stops at $\geq$0.90. Only the first 7 of up to 20 steps are shown: a row ending in … continues, and $|S|$ is the full length of its sequence. Items are FAQ unless prefixed (CDT/CCT = clock drawing/copying). exact age+sex matched pairs, every pair the split affords; greedy search run on the full cohort rather than a 40-patient subsample.

Model	Seed	n	step 1	step 2	step 3	step 4	step 5	step 6	step 7	$\|S\|$	final
AD→CN — normalising edits, patients called AD											
6L d256 clf	0	303	Taxes	Appointments	Checks	Games	Travel	CDT:hands_time	Attention …	20	0.86
6L d256 clf	2	298	Appointments	Taxes	Games	Checks	Travel	Attention	Shopping …	20	0.87
6L d256 gen	0	295	Appointments	Taxes	Checks	Games	Travel	Attention	Shopping …	11	0.92
6L d256 gen	2	313	Appointments	Taxes	Checks	Travel	Games	Attention	CurrentEvents …	10	0.92
12L d256 clf	0	308	Appointments	Taxes	Checks	Travel	Games	Attention	CurrentEvents …	10	1.00
12L d256 clf	2	313	Appointments	Taxes	Travel	Games	Checks	Attention	CurrentEvents …	10	0.99
12L d256 gen	0	310	Appointments	Taxes	Checks	Games	Travel	Attention	Shopping …	11	0.92
12L d256 gen	2	310	Appointments	Checks	Taxes	Games	Attention	Travel	Meals(bal) …	14	0.91
12L d768 clf	0	313	Appointments	Taxes	Checks	Games	Travel	Attention	Shopping …	12	0.90
12L d768 clf	2	315	Appointments	Taxes	Checks	Games	Travel	Meals(bal)	Attention …	10	0.98
12L d768 gen	0	317	CDT:hands_time	Checks	Taxes	Shopping	Games	Meals(hot)	Meals(bal) …	20	0.01
12L d768 gen	2	308	Appointments	Taxes	Checks	Games	Travel	Attention	CurrentEvents …	10	1.00
1.4B clf	0	301	Taxes	Checks	Meals(bal)	Appointments	Travel	Games	Shopping …	20	0.06
1.4B clf	2	304	Taxes	Appointments	Checks	Travel	Games	Attention	Shopping …	9	0.92
1.4B gen	0	266	Taxes	Appointments	Checks	Games	Meals(bal)	Shopping	Attention …	9	0.96
1.4B gen	2	301	Taxes	Appointments	Games	Meals(bal)	Shopping	Travel	Attention …	20	0.23
CN→AD — impairing edits, patients called CN											
6L d256 clf	0	303	Appointments	Taxes	–	–	–	–	–	2	1.00
6L d256 clf	2	298	Shopping	–	–	–	–	–	–	1	1.00
6L d256 gen	0	295	Checks	–	–	–	–	–	–	1	1.00
6L d256 gen	2	313	Taxes	–	–	–	–	–	–	1	1.00
12L d256 clf	0	308	Checks	–	–	–	–	–	–	1	1.00
12L d256 clf	2	313	Checks	–	–	–	–	–	–	1	1.00
12L d256 gen	0	310	Appointments	–	–	–	–	–	–	1	0.90
12L d256 gen	2	310	Attention	–	–	–	–	–	–	1	0.98
12L d768 clf	0	313	Checks	Meals(hot)	Travel	–	–	–	–	3	0.99
12L d768 clf	2	315	Checks	–	–	–	–	–	–	1	1.00
12L d768 gen	0	317	CDT:hands_time	CDT:two_hands	Appointments	Meals(hot)	Games	–	–	5	0.95
12L d768 gen	2	308	Appointments	Meals(bal)	–	–	–	–	–	2	1.00
1.4B clf	0	301	Travel	Appointments	Shopping	CDT:numbers	CCT:hands_time	CDT:hands_time	Meals(hot) …	20	0.55
1.4B clf	2	304	Travel	–	–	–	–	–	–	1	0.94
1.4B gen	0	266	Attention	CurrentEvents	Meals(hot)	CDT:circular_face	Meals(bal)	CDT:symmetry	CDT:numbers …	8	0.91
1.4B gen	2	301	Attention	Meals(hot)	Appointments	Taxes	Shopping	–	–	5	0.90

Table 22: Is the minimal sufficient set also necessary? Items are accumulated greedily, every remaining candidate re-scored at each step; the selected sequences are listed in the greedy-sequence table. Search, stopping rule and every rate reported here are evaluated on the SAME full cohort of n pairs, so *reached* says whether flip(S) met the 0.90 threshold on those pairs rather than on a subsample. *necessity* is the range of flip(S)−flip($S \setminus s$) over elements s. *random* is the mean flip of size-matched random subsets and p_{random} the one-sided permutation p, $(1 + \#\{\mathrm{random} \geq \mathrm{flip}(S)\})/(1 + R)$ over R draws – the control separating *these* $|S|$ items from *this many* items. *subst* is the best flip after swapping one element for an item outside S; recovering flip(S) means the element is necessary without being special. p_{subst} tests that swap: an exact one-sided McNemar of S against its STRONGEST substitution on the same pairs, so a small value means S beats every one-element swap and the elements are special rather than class-substitutable. The two p-values are not comparable and answer different questions – cardinality for p_{random}, identity for p_{subst} – and neither is corrected for multiplicity, the substitution test because comparing against the strongest competitor is already the conservative direction. † greedy never reached the stop threshold even using all 20 items, so there is no sufficient set and necessity within it is undefined; the value shown is the ceiling reached (6L d256 clf s0 AD2CN, 6L d256 clf s2 AD2CN, 12L d768 gen s0 AD2CN, 1.4B clf s0 AD2CN, 1.4B gen s2 AD2CN, 1.4B clf s0 CN2AD). Cohort: exact age+sex matched pairs, every pair the split affords; greedy search run on the full cohort rather than a 40-patient subsample.

Model	Seed	n	$\lvert S\rvert$	flip(S)	reached	necessity	random	p_{random}	subst	p_{subst}
AD→CN — normalising edits, patients called AD										
6L d256 clf	0	303	–	0.86†	no	–	–	–	–	–
6L d256 clf	2	298	–	0.87†	no	–	–	–	–	–
6L d256 gen	0	295	11	0.92	yes	0.02–0.65	0.17	<0.05	0.91	0.23
6L d256 gen	2	313	10	0.92	yes	0.12–0.66	0.12	<0.05	0.81	<0.01
12L d256 clf	0	308	10	1.00	yes	0.13–0.72	0.11	<0.05	0.87	<0.01
12L d256 clf	2	313	10	0.99	yes	0.13–0.70	0.12	<0.05	0.87	<0.01
12L d256 gen	0	310	11	0.92	yes	0.05–0.65	0.17	<0.05	0.88	<0.01
12L d256 gen	2	310	14	0.91	yes	0.02–0.60	0.29	<0.05	0.90	0.06
12L d768 clf	0	313	12	0.90	yes	0.07–0.58	0.17	<0.05	0.85	<0.01
12L d768 clf	2	315	10	0.98	yes	0.12–0.68	0.12	<0.05	0.87	<0.01
12L d768 gen	0	317	–	0.01†	no	–	–	–	–	–
12L d768 gen	2	308	10	1.00	yes	0.10–0.66	0.14	<0.05	0.90	<0.01
1.4B clf	0	301	–	0.06†	no	–	–	–	–	–
1.4B clf	2	304	9	0.92	yes	0.24–0.61	0.12	<0.05	0.72	<0.01
1.4B gen	0	266	9	0.96	yes	0.06–0.14	0.28	<0.05	0.97	0.81
1.4B gen	2	301	–	0.23†	no	–	–	–	–	–
CN→AD — impairing edits, patients called CN										
6L d256 clf	0	303	2	1.00	yes	0.23–0.46	0.49	0.10	1.00	1.00
6L d256 clf	2	298	1	1.00	yes	1.00–1.00	0.32	0.10	0.95	<0.01
6L d256 gen	0	295	1	1.00	yes	1.00–1.00	0.50	0.29	1.00	1.00
6L d256 gen	2	313	1	1.00	yes	1.00–1.00	0.22	0.19	1.00	1.00
12L d256 clf	0	308	1	1.00	yes	1.00–1.00	0.30	0.33	1.00	1.00
12L d256 clf	2	313	1	1.00	yes	1.00–1.00	0.30	0.33	1.00	1.00
12L d256 gen	0	310	1	0.90	yes	0.90–0.90	0.28	<0.05	0.89	0.09
12L d256 gen	2	310	1	0.98	yes	0.98–0.98	0.22	0.10	0.75	<0.01
12L d768 clf	0	313	3	0.99	yes	0.13–0.28	0.66	<0.05	1.00	1.00
12L d768 clf	2	315	1	1.00	yes	1.00–1.00	0.34	0.33	1.00	1.00
12L d768 gen	0	317	5	0.95	yes	0.01–0.82	0.16	<0.05	1.00	1.00
12L d768 gen	2	308	2	1.00	yes	0.89–0.90	0.17	0.14	1.00	1.00
1.4B clf	0	301	–	0.55†	no	–	–	–	–	–
1.4B clf	2	304	1	0.94	yes	0.94–0.94	0.04	<0.05	0.21	<0.01
1.4B gen	0	266	8	0.91	yes	0.06–0.09	0.58	<0.05	0.91	1.00
1.4B gen	2	301	5	0.90	yes	0.00–0.02	0.76	<0.05	0.90	0.75

Sets robust to one-item substitution. Among the sufficient sets above, we additionally identify the runs for which the selected set significantly outperforms its strongest tested one-item substitution. Table 23 reports this filtered subset.

Table 23: **Greedy sufficient edit sets with stronger effects than their best tested one-item substitution.** Shown are the 11 of 26 sufficient-set runs for which S significantly outperforms its strongest tested one-element substitution under an exact one-sided McNemar test ($p < 0.05$); all remain significant after Benjamini–Hochberg correction across the 26 comparisons. *subst* gives the flip rate of the best substitution, and $|S|$ the full set size. Only the first three greedily selected items are shown.

Condition	$\|S\|$	flip(S)	subst	p_{subst}	Selected sequence
AD→CN — normalising features					
6L *d*256 gen (s2)	10	0.92	0.81	$< 10^{-7}$	Appointments → Taxes → Checks → …
12L *d*256 clf (s0)	10	1.00	0.87	$< 10^{-11}$	Appointments → Taxes → Checks → …
12L *d*256 clf (s2)	10	0.99	0.87	$< 10^{-12}$	Appointments → Taxes → Travel → …
12L *d*256 gen (s0)	11	0.92	0.88	$< 10^{-2}$	Appointments → Taxes → Checks → …
12L *d*768 clf (s0)	12	0.90	0.85	$< 10^{-3}$	Appointments → Taxes → Checks → …
12L *d*768 clf (s2)	10	0.98	0.87	$< 10^{-9}$	Appointments → Taxes → Checks → …
12L *d*768 gen (s2)	10	1.00	0.90	$< 10^{-9}$	Appointments → Taxes → Checks → …
1.4B clf (s2)	9	0.92	0.72	$< 10^{-11}$	Taxes → Appointments → Checks → …
CN→AD — impairing features					
6L *d*256 clf (s2)	1	1.00	0.95	$< 10^{-3}$	Shopping
12L *d*256 gen (s2)	1	0.98	0.75	$< 10^{-21}$	Attention
1.4B clf (s2)	1	0.94	0.21	$< 10^{-66}$	Travel

D.7 Analytical Limitations and Scope

Training regime does not isolate pretraining alone. Our architecture-matched comparison does not identify a pure causal effect of pretraining. Pythia combines language-model pretraining with LoRA adaptation, whereas the GPT-NeoX controls are randomly initialised and trained directly on the downstream task; the two branches also differ in parameter adaptation and learning-rate selection. We therefore interpret differences between them as effects associated with the complete *training regime*, rather than attributing them uniquely to pretraining. The comparison remains informative because the matched pairs hold fixed the decoder-only model family and the principal architectural dimensions—depth, hidden width, and attention-head count—thereby removing several major architectural confounds.

The endpoint is broader than Alzheimer's disease alone. The AD/CN notation used throughout the experiments follows the repository's binary output labels, but the positive class is a clinically impaired cohort containing Alzheimer's disease together with early and late mild cognitive impairment, while the control class contains cognitively normal participants and a small number with subjective memory impairment. The results should therefore be interpreted as mechanisms of this ADNI-derived cognitive-status classification task rather than as mechanisms specific to Alzheimer's disease. This does not affect the within-model comparisons that form the main claims, because all models are evaluated on the same underlying task definition, but it does limit disease-specific clinical interpretation.

Mechanistic analyses reuse the IID test cohort. Several mechanistic analyses are performed on subsets of the same IID test cohort used for predictive evaluation. Most notably, circuit extraction ranks heads on a circuit-fit subset and evaluates faithfulness on a disjoint circuit-held-out subset, although both are drawn from the official IID test partition. The held-out subset therefore tests whether the extracted circuit generalises beyond the examples used for head selection, but it is not an untouched mechanistic test set. This design remains appropriate for the paper's primary aim of comparing models under a common intervention protocol, since the sampling, extraction, and faithfulness criteria are held fixed across conditions. A stricter validation would perform circuit selection on the training or validation partition and reserve an untouched IID test subset for mechanistic evaluation. External-cohort evaluation would address a different question, namely mechanistic robustness under distribution shift, and would first require characterising the shift and its effects on predictive performance, margins, and intervention eligibility before differences in mechanism could be interpreted.

Readout patching captures direct readout-mediated influence. Activation patching is performed only at the task-specific readout position, so it measures the direct downstream effect of replacing the state available there at a given layer. Information written at other token positions and used later through attention may therefore be missed, and the layer with the largest patching effect need not be the layer at which diagnostic information first becomes available. Classifier and generative models also use different readout contexts: the final narrative position and the `Diagnosis:` cue, respectively; so their trajectories compare complete task formulations rather than isolating the output head.

These constraints still permit a controlled comparison of where readout-mediated diagnostic influence is strongest: the same patching procedure is applied across matched models and both donor directions, and final-residual patching provides a positive control for intervention efficacy. We therefore interpret the results as differences in the depth profile of direct readout-mediated influence, not as a complete localisation of all diagnostic computation. Future work could trace information flow across multiple token positions and component pathways.

Counterfactual rewrites test score invariance rather than a unique linguistic mechanism. The input interventions are controlled synthetic perturbations of the clinical narrative rather than observed changes in a patient's clinical state. In particular, `Never did, but could do now (0)` preserves the FAQ score relative to `Normal (0)`, but it is not semantically identical because it also changes the reported history of performing the activity. The resulting flips therefore provide direct evidence that model predictions are not invariant to score-preserving wording, but they do not by themselves identify lexical overlap as the sole cause.

The length-matched neutral-filler control helps narrow this interpretation: it produces little or no effect in most evaluated runs, arguing against sequence length alone as a general explanation, although one Pythia-1.4B generative solution is itself sensitive to the filler. This does not weaken the primary conclusion, which concerns failure of score invariance rather than a specific linguistic mechanism. More complete separation of lexical, syntactic, pragmatic, and distributional effects would require a broader set of independently generated score-preserving paraphrases and additional clinical-language controls.

Sufficiency and necessity describe intervention-specific decision rules. Feature sufficiency and necessity are defined relative to particular edits, directions, and model-specific eligible cohorts. A high sufficiency value therefore means that the tested rewrite can change the model's prediction in that context; it does not imply that the corresponding clinical feature is unconditionally sufficient for diagnosis. Likewise, leave-one-out necessity measures how much of a joint-intervention effect is lost when one feature is withheld, rather than providing a global feature-importance score.

This conditionality is also informative. In settings where single-item sufficiency saturates, necessity can still distinguish which features are more important to the joint intervention. Reporting both quantities therefore allows us to compare how models organise evidence under controlled perturbations without conflating model dependence with clinical importance. Broader claims about the clinical significance of these features require evidence external to the model interventions themselves.

Extracted subsets are conditional on the search and intervention procedure. Both the attention-circuit and clinical-feature analyses identify sufficient subsets under a specified intervention and greedy search procedure rather than globally minimal or unique explanations. For attention circuits, heads are ranked by their ablation effect and restored until 90% two-point faithfulness is recovered, while embeddings, MLPs, and the task readout remain intact. For clinical feature sets, every remaining feature is re-scored conditional on the currently selected set and added greedily until the 90% flip-rate target is reached or all features are exhausted. In both cases, interactions or redundancy could permit alternative subsets that the search does not visit.

The comparisons remain informative because each analysis applies the same search rule, intervention definition, and stopping criterion across matched models. We therefore interpret the results as relative attention-head compressibility and greedily identified sufficient feature sets under common protocols, rather than as unique global circuits or decision rules. For the feature analysis, size-matched random subsets and one-item substitutions provide additional evidence that some solutions

depend on feature identity rather than cardinality alone, although they do not exhaust the combinatorial search space. Future work could test robustness to alternative ablation baselines, search algorithms, and more exhaustive subset optimisation.

Cross-model intervention effects are conditional on eligible cohorts and analysed runs. Several interventions are evaluated only on participants that the corresponding model classifies correctly; readout patching additionally requires sufficient baseline-margin separation to keep the normalised effect well defined. These gates are necessary for interpretable intervention estimates, but they make the eligible cohort model-specific. Cross-model differences in flip rates or patching magnitude should therefore be read as comparisons of within-model response profiles under a common protocol, rather than as paired causal effects of changing scale, training history, or task formulation on an identical set of predictions.

We reduce this source of variation by using fixed data splits, the same matching criteria and intervention definitions across models, and both intervention directions. Nevertheless, model-specific eligibility remains part of the estimand. In addition, although predictive evaluation and selected circuit summaries use five training seeds, the full mechanistic and counterfactual pipeline is carried out on two analysed runs (seed 0 and seed 2). The substantial differences observed between these runs are themselves evidence that mechanistic properties can vary across independently trained solutions, but they limit claims about how frequently each pattern occurs. Future work should extend the complete intervention pipeline across more training runs and, where possible, report both model-specific and common-eligible-cohort analyses.

Demographic swapping is a mechanism control, not a fairness audit. The near-null effects of sex, recorded race, and year-of-birth activation swaps should not be interpreted as evidence that the models are demographically unbiased. The matching procedure deliberately conditions on cognitive and functional similarity, which changes the evaluated cohort and can underrepresent more clinically distinct cases. Moreover, the intervention tests a narrow mechanistic question: whether replacing the selected demographic representation changes the model's prediction.

This design is nevertheless useful as a negative mechanism control. The same models show much larger responses to functional-feature interventions, and the final-residual positive control confirms that the patching procedure is capable of moving predictions when applied at a causally effective site. Thus, the near-null demographic effects provide evidence of limited direct readout-mediated influence under the tested intervention, not evidence of fairness or absence of demographic bias. A dedicated fairness analysis would require appropriately powered subgroup evaluation and methods designed to assess demographic performance disparities.

Causal claims are within-model and external validity is deliberately scoped. Throughout this work, "causal" refers to interventions on model inputs or internal activations and the resulting changes in the model's own prediction. These interventions provide stronger evidence of model dependence than correlation, probing, or attribution alone, but they do not estimate causal effects in patients, establish disease aetiology, or demonstrate a single mediated pathway from the identified internal components to the observed feature-level effects.

The study is deliberately restricted to one ADNI-derived task, a common decoder-only architectural family, and structured template-generated narratives. This controlled setting enables architecture-matched comparisons and targeted manipulation of numerical and verbal clinical features, reducing several sources of variation that would complicate mechanistic comparison. The corresponding trade-off is external validity: the observed mechanisms should be interpreted as properties of the evaluated models, cohort, and interventions rather than as general properties of clinical language models. Future work should test whether these patterns replicate across independent clinical cohorts, naturally occurring clinical text, and additional model families.