Where Does Mathematical Skill Live? A Controlled Biopsy of Transformer Mathematical Representations

XRayBench v0.2: Seven controlled phases and the Partial Structure Index assess option-scoring signals in transformer models

Dr. Tamás Nagy

ORCID: 0009-0004-8079-4679

Independent Researcher

tnagyphd@gmail.com

https://thalens.org

Working Paper • 2026-08-20

Build-std v2.0 | 2026-08-20 18:32 | 50ef87d3

Overview

Language models appear to solve mathematical problems, but where does this ability reside inside the network? A growing body of interpretability research uses layer-by-layer option scoring to claim that transformers develop internal mathematical representations. These claims are rarely tested against rigorous controls.

We introduce XRayBench, a seven-phase controlled benchmark that assesses option-scoring signals against three confounds: token-frequency bias, token-length bias, and surface-pattern matching. Full XRay Scores are reported for three GPT-2-family checkpoints, while PSI-only measurements extend to GPT-2-large and four Pythia checkpoints. In the three full runs, the reported controls match or exceed the real-task option-scoring signal at the aggregate level.

However, the Partial Structure Index (PSI) — a per-task metric that subtracts the stronger of two option-scoring controls from the real signal — shows a family-dependent pattern in these measurements: **logic-requiring families** (variable binding, theorem patterns, proof closure) have higher PSI than **computation families** (arithmetic, algebra) at several tested scales. A descriptive threshold rule reaches 98% accuracy because it uses a direct component of PSI. The pattern is examined on held-out parameterizations with no exact prompt overlap and in a preliminary Pythia replication under a first-token proxy, where the Logic−Compute PSI gap is positive at all tested scales and reaches +0.28 at Pythia-1B.

The contribution is both methodological (a reusable controlled protocol) and empirical (controlled option-scoring evidence of family-dependent patterns in the tested transformer models).

Abstract

We present XRayBench v0.2, a seven-phase controlled benchmark for evaluating option-scoring evidence about mathematical representations in transformer language models. The executed training

battery contains 50 tasks in 5 mathematical families, then evaluates each signal against three control conditions (random-label, matched-token, semantics-breaking), paraphrase stability, activation patching, and MLP/attention ablation; ablation phases are implemented but their results are not reported in v0.2. We define the XRay Score, a single metric in $[-4, +4]$ summarizing reported evidence phases, and introduce the **Partial Structure Index (PSI)**, a per-task metric for residual option-scoring signal beyond the random-label and matched-token controls.

For distilgpt2 (6 blocks, 82M), GPT-2 (12 blocks, 117M), and GPT-2-medium (24 blocks, 355M), XRay Scores are non-positive. Confound dominance varies by checkpoint, with matched-token controls producing the largest gap for GPT-2. PSI-only analyses additionally cover GPT-2-large (36 blocks, 774M) and Pythia models from 70M to 1B parameters. In these measurements, **PSI shows a family-dependent pattern**: the unweighted mean over binder_tracking, theorem_pattern, and proof_closure is higher, but non-monotone, than the unweighted mean over arithmetic and algebra at several tested scales. A descriptive threshold rule reaches 98% accuracy using n_real_win_layers, a direct PSI component. The pattern is examined on held-out parameterizations and preliminarily cross-checked on Pythia under a different, first-token protocol.

The global XRay Score masks a family-dependent pattern in the tested measurements: **logic and computation show different PSI profiles over the evaluated models.**

Keywords: mechanistic interpretability, mathematical reasoning, controlled benchmark, logit lens, token bias, representation probe, partial structure index, emergent representations

1. Introduction

1.1 The Problem

The "logit lens" technique and its extensions enable reading a transformer's predictions at intermediate layers by projecting hidden states through the unembedding matrix. Applied to mathematical tasks, this reveals intriguing patterns: models sometimes "know" the correct answer at middle layers before losing it at the output. These observations have fueled claims about dark mathematical skill — internal representations of mathematical knowledge that exist even when the model's final output is wrong.

The core problem is that these observations are uncontrolled. A language model that has memorized token co-occurrence statistics will also show preferential activation for certain options at intermediate layers. Without baselines, "the correct answer wins at layer 7" is a statement about token priors, not mathematical reasoning.

1.2 The Controlled Benchmark Approach

We propose that any claim about internal mathematical skill must survive three falsification tests:

Option-scoring controls

The three controls intervene on different parts of the option-scoring setup and therefore support different conclusions.

Control	Intervention	Interpretive limit
Random-label	Relabel a wrong option as "correct"	Persisting wins challenge an interpretation based only on mathematical correctness.
Matched-token	Relabel an exact-token-length wrong option when available, otherwise the nearest-length wrong option	Reduces one length-related difference without isolating content fully.
Semantics-breaking	Offset prompt numbers or replace question tails with nonsense	Challenges a purely semantic interpretation; does not isolate semantics.

Additional robustness criteria

Additionally, a mathematical-representation hypothesis should predict:

- **Prompt-stable:** the same computational layers respond to the mathematical content regardless of how the question is phrased (measured by paraphrase Jaccard similarity).
- **Causally active:** ablating specific layers should differentially affect mathematical tasks versus controls.
- **Geometrically coherent:** hidden-state representations of the same mathematical content should cluster, independent of option scoring.

XRayBench specifies all six tests in a single automated protocol; v0.2 reports results for the option-scoring, patching, and paraphrase phases, but not for the MLP/attention ablation phase.

1.3 Main Results

Finding 1 (Global). In the three GPT-2-family checkpoints with complete XRay runs (82M–355M parameters), the XRay Score is non-positive. In these option-scoring measurements, the reported controls match or exceed the real-task aggregate signal.

Finding 2 (Family-Dependent Pattern). The Partial Structure Index (PSI) reveals a family-dependent contrast hidden by the global score. Three small logic-labelled families — binder_tracking, theorem_pattern, and proof_closure — differ from the two more numerous computation-labelled families:

Family	distilgpt2 (6L)	GPT-2 (12L)	GPT-2-medium (24L)
binder_tracking	−0.19	+0.14	+0.27
theorem_pattern	−0.22	−0.21	+0.08
proof_closure	−0.44	−0.43	+0.13
arithmetic_transform	−0.22	−0.35	−0.37
algebra_simplification	−0.33	−0.21	−0.33

Computation families (arithmetic, algebra) remain negative at all four GPT-2-family scales. Using unweighted means of family means, the Logic−Compute PSI gap is positive but non-monotone, moving from approximately zero at 82M to +0.51 at 355M and +0.30 at 774M. A descriptive threshold rule reaches 98% accuracy on training tasks

and **86% on held-out parameterizations of the same task templates with no exact prompt overlap**; because its primary feature is a direct PSI component, this does not establish an independently learnable or generalizable boundary.

Finding 3 (Cross-Architecture Replication). The dichotomy is not GPT-2-specific. On Pythia-70M, 160M, 410M, and 1B, the Logic−Compute PSI gap is positive at every tested scale:

Model	Logic PSI	Compute PSI	Logic−Compute Gap
Pythia-70M	−0.1139	−0.1328	+0.0189
Pythia-160M	−0.1409	−0.2336	+0.0927
Pythia-410M	−0.0224	−0.2520	+0.2295
Pythia-1B	+0.0525	−0.2303	+0.2828

The gap is positive at all four Pythia scale points and increases within that first-token-proxy series. Because the GPT-2 and Pythia protocols differ, we do not pool them into one scaling estimate.

1.4 Comparison with Prior Work

Method	Controls	Paraphrase test	Ablation	Multi-model	Score metric
Logit	No	No	No	No	No
Tuned	No	No	No	Yes	No
Tracing	Partial	No	Yes	No	No
Probing	Yes	No	No	Yes	Accuracy
XRayBench (this work)	**3 types**	**Yes**	**Implemented; not reported in v0.2**	**Yes**	**XRay Score**

Note: Logit, Tuned, Tracing, and Probing abbreviate the logit lens, tuned lens, causal tracing, and probing-classifier methods. The table records this paper's classification of *interpretability methods* for reading internal representations; sources are nostalgebraist (2020), Belrose et al. (2023), Meng et al. (2022), and Belinkov (2022). Behavioral evaluations establish mathematical capabilities at the model-output level (Saxton et al., 2019; Lewkowycz et al., 2022). More directly, Stolfo et al. (2023) causally localize arithmetic processing and Hou et al. (2023) use attention-based probes to recover intermediate reasoning structure. These are direct precedents for studying internal mathematical computation. XRayBench's narrower contribution is a packaged comparison against a fixed battery of option-scoring controls, paraphrases, held-out parameterizations, and multiple model families, summarized by the PSI and XRay metrics. Razeghi et al. (2022) further show why frequency controls matter: few-shot numerical performance can track pretraining term frequencies.

1.5 Organization

Section 2 defines the seven-phase protocol. Section 3 describes the task battery. Section 4 presents global results across GPT-2-family models. Section 5 introduces the Partial Structure Index and

the scaling dichotomy. Section 6 validates the findings with held-out tasks, machine-extracted conjectures, and Pythia cross-architecture replication. Section 7 analyzes the theorem_transitive exception with orthogonal representation probes. Section 8 discusses implications and limitations.

2. The XRayBench Protocol

2.1 Layer-by-Layer Option Scoring

For a prompt p with answer options $\{o_1, \ldots, o_k\}$, we compute multi-token log-probability scores at each layer ℓ:

$$S_\ell(o_i) = \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \log P_\ell(o_{i,t} \mid p, o_{i,<t})$$

where P_ℓ is obtained by projecting the layer-ℓ residual stream through layer normalization and the unembedding matrix, then applying softmax. Scores are prior-normalized by subtracting $S_\ell^{\text{neutral}}(o_i)$ computed under a content-free prompt ("Answer:").

The **internal win** at layer ℓ is 1 if $\arg\max_i S_\ell(o_i)$ equals the correct option.

2.2 Seven Phases

Building on the position-level score, the protocol adds controls and robustness checks in a fixed sequence.

Phase	Input	What it measures
1	50 real mathematical tasks	Baseline internal win rate
2	Same tasks, random option relabeled as correct	Label-assignment independence
3	Same tasks, exact- or nearest-token-length wrong option relabeled	Token-length sensitivity
4	Same tasks, mathematical content broken	Content necessity
5	Known-then-lost tasks, carry-forward patch	Causal signal from early correct layers
6	8 task groups × 5 paraphrase variants	Prompt robustness (Jaccard)
7	MLP and attention ablation sweep	Computational layer identification

2.3 The XRay Score

The phases produce heterogeneous evidence, so the XRay Score compresses four reported components into one signed summary.

We define a composite evidence metric:

$$\text{XRay Score} \in \{-4, -3, \ldots, +3, +4\}$$

with one point awarded for each of four conditions:

Condition	+1 if	0 if	−1 if
vs Random-label	real_win > rl_win	equal	real_win < rl_win
vs Matched-token	real_win > mt_win	equal	real_win < mt_win
vs Semantics-break	real_win > sb_win	equal	real_win < sb_win
Paraphrase stability	$\overline{J} > 0.5$	$\overline{J} = 0.5$	$\overline{J} < 0.5$

A positive XRay Score means real mathematical tasks outperform the reported controls in this probe. A negative score means at least some controls match or exceed the observed real-task signal; neither sign alone establishes or excludes mathematical skill in the model.

3. Task Battery

3.1 Executed Five-Family Battery

The implementation constructs a larger ordered task pool and the reported runs execute its first 50 entries. The resulting battery is imbalanced: 38 tasks are arithmetic or algebra variants, while each of the three logic-labelled families has 4 tasks. The executed population is:

Family	Tasks	Cognitive layer	Example
arithmetic_transform	24	Numerical transformation	"Start with 7. Add 5. The result is"
algebra_simplification	14	Symbolic transformation	"Simplify x minus x. The result is"
binder_tracking	4	Binding/reference	"Let a = 3, b = a + 2. Then b equals"
theorem_pattern	4	Relation pattern	"If A implies B, B implies C, then A implies"
proof_closure	4	Proof-state completion	"If a proof has both P and not P, it closes by"

The family labels are author-defined benchmark categories. Family means for the three 4-task groups should be interpreted as pilot estimates, not population-level measurements.

3.2 Control Generation

Random-label controls randomly reassign which option is labeled "correct." **Matched-token controls** choose a wrong option with the same tokenized length as the real correct option when

one exists; otherwise they choose from the closest-length wrong options. This reduces one length-related difference but does not guarantee an exact match or isolate all content effects. **Semantics-breaking controls** perturb the prompt while leaving the answer options unchanged: arithmetic and binder prompts have each integer shifted independently by 2–5, while other task families receive a grammatical-nonsense question tail (for example, "the color is" or "the planet is"). These perturbations do not preserve every surface statistic and therefore do not isolate semantics on their own.

3.3 Paraphrase Groups

Eight tasks are expanded into 5 paraphrase variants each. For example, "Start with 7. Add 5. The result is" becomes {"7 plus 5 equals", "What is 7 + 5? The answer is", "Begin with 7 and add 5. You get", "Calculate 7 + 5. The result is"}. The Jaccard similarity of win-layers across paraphrases measures prompt invariance.

4. Results

4.1 XRay Score Across Scale

Model	Layers	Params	Real Win	RL Ctrl	MT Ctrl	SB Ctrl	Jaccard	XRay Score
distilgpt2	6	82M	0.50	0.66	0.62	0.48	0.64	**0**
GPT-2	12	117M	0.46	0.68	0.78	0.52	0.53	**−2**
GPT-2-medium	24	355M	0.64	0.80	0.80	0.68	0.46	**−4**

The XRay Score decreases with model size in the three models tested (though N=3 is insufficient for statistical significance). This does not mean larger models have less mathematical skill; it means the option-scoring probe increasingly measures token statistics rather than content at larger scales.

4.2 Gap Analysis

Model	Gap vs RL	Gap vs MT	Gap vs SB
distilgpt2	−0.16	−0.12	+0.02
GPT-2	−0.22	−0.32	−0.06
GPT-2-medium	−0.16	−0.16	−0.04

All reported point-estimate gaps are negative or near-zero; no significance test is claimed. Confound dominance varies by checkpoint: random-label has the larger gap for distilgpt2, random-label and matched-token tie for GPT-2-medium, and matched-token is largest for GPT-2. For GPT-2, the −0.32 matched-token gap means the fake-correct option selected by the exact-or-nearest token-length rule wins at some internal position on 32 percentage points more tasks than the real correct option.

4.3 Carry-Forward Patching

Model	Flip Rate	Mean Best Delta
distilgpt2	0.59	+2.06
GPT-2	0.57	+2.54
GPT-2-medium	0.60	+3.47

Injecting an earlier layer's residual state into a later layer does improve final-layer discrimination (flip rate ~60%, i.e., more than half of "known-then-lost" trajectories can be recovered). However, the flip rate does not distinguish real from control tasks. This is consistent with token-level information being the recovered signal.

4.4 Paraphrase Stability

Model	Mean Jaccard	theorem_transitive_01 Jaccard
distilgpt2	0.64	0.17
GPT-2	0.53	0.20
GPT-2-medium	0.46	0.83

Global paraphrase stability decreases with scale, while the single theorem_transitive_01 prompt group increases from 0.17 to 0.83 across these three checkpoints. Each group contains five paraphrases of one base task; it is not a three-task subset. Several other groups attain Jaccard 1.0 because every paraphrase follows the same never_known trajectory, so a high Jaccard value alone is not evidence of mathematical structure. The transitivity group is discussed further only as a descriptive exception that combines stability with nonzero internal and final accuracy.

5. The Partial Structure Index

5.1 Motivation

The XRay Score aggregates evidence across all task families. This is appropriate for the global option-scoring question but masks family-specific patterns. We introduce the **Partial Structure Index (PSI)** to measure per-task structure not explained by two option-scoring controls.

5.2 Definition

For each task t and model M, we run three trajectory panels (real, random-label control, matched-token control). Let L_M be the recorded positions: the embedding, every zero-indexed transformer-block output, and the final output. Let $W_c(t, M) \subseteq L_M$ be the positions where the option labelled correct wins under condition c. PSI is defined as:

$$\mathrm{PSI}(t, M) = J_{\mathrm{real}}(t, M) - \max\left(J_{\mathrm{rl}}(t, M),\ J_{\mathrm{mt}}(t, M)\right)$$

where $J_c(t, M) = |W_c(t, M)|/|L_M|$, equivalently the Jaccard similarity of W_c with the full recorded-position set. Thus J_{real}, J_{rl}, and J_{mt} use the same model-specific denominator.

PSI > 0 means the real task has a larger fraction of correct-winning recorded positions than either reported option-scoring control. **PSI 0** means at least one of those controls matches or exceeds that fraction. PSI does not by itself exclude semantics-breaking, paraphrase, or ablation confounds.

We also define an extended variant $\text{PSI}_{\text{ext}} = \text{PSI} + 0.3 \cdot \sigma$, where $\sigma \in \{-1, 0, +1\}$ is a stability bonus: $+1$ if the real task has stable winning spans that controls lack, -1 if controls have stable spans that the real task lacks.

5.3 The Scaling Dichotomy

PSI by family across model scale reveals a striking split:

Family	distilgpt2 (6L)	GPT-2 (12L)	GPT-2-medium (24L)	GPT-2-large (36L)
binder_tracking	−0.19	+0.14	**+0.27**	+0.16
theorem_pattern	−0.22	−0.21	+0.08	−0.04
proof_closure	−0.44	−0.43	**+0.13**	−0.09
arithmetic_transform	−0.22	−0.35	−0.37	−0.27
algebra_simplification	−0.33	−0.21	−0.33	−0.31

Logic-requiring families cross from negative to positive between GPT-2 (12L) and GPT-2-medium (24L), with binder_tracking reaching +0.27 — the strongest individual PSI in any family. At GPT-2-large (36L), logic families remain systematically higher than computation families, though the PSI peak for most families occurs at GPT-2-medium. Computation families remain negative at all four scales.

This is not visible in the global XRay Score, which averages over all families and is dominated by the more numerous computation tasks.

For a compact descriptive contrast, we take an unweighted mean of the three logic-labelled family means and an unweighted mean of the two computation-labelled family means:

Group	distilgpt2 (82M)	GPT-2 (117M)	GPT-2-medium (355M)	GPT-2-large (774M)
Logic families	−0.2813	−0.1667	**+0.1602**	+0.0110
Computation families	−0.2772	−0.2843	−0.3483	−0.2888
L−C gap	**−0.004**	**+0.118**	**+0.509**	**+0.300**

At distilgpt2, the gap is near zero. It is +0.51 at GPT-2-medium and +0.30 at GPT-2-large. Thus, the tested GPT-2 models show a positive but non-monotone Logic−Compute gap; the present data do not identify a critical window for logic-family emergence.

5.4 Mutual Information Analysis

We compute mutual information between task features and positive PSI across all models:

Rank	Feature	MI(feature; PSI > 0)
1	n_real_win_layers	0.4577
2	stability_bonus	0.4354
3	j_real	0.3485
4	j_mt	0.0894
5	family	0.0620

The number of correct-winning recorded positions in the real task has the highest measured MI, followed by the stability bonus. Because these quantities directly enter PSI or its extension, this is a descriptive property of the metric rather than evidence for an independent predictor.

5.5 Decision Tree

A depth-3 decision tree classifies PSI > 0 with **98% accuracy** (n=150, 3 models $\times$ 50 tasks). Its primary split is on n_real_win_layers, which is a direct component of PSI; the tree is therefore a descriptive threshold rule, not an independent predictor of mathematical structure.

This 98% accuracy reflects the metric's construction and the observed task measurements. Section 6 reports how the same descriptive rule behaves on held-out parameterizations.

6. Validation: Held-Out Tasks and Conjecture Testing

6.1 The Held-Out Battery

To test sensitivity to non-identical parameterizations, we generate a held-out set of 54 mathematical tasks from the same broad templates:

- New arithmetic operand combinations, with some individual values reused
- Additional variable-name combinations, with some single-letter tokens reused
- Additional relational contexts, including new implication chains
- Additional proof-closure concepts, including introduction and case analysis

No held-out prompt is exactly identical to a training prompt. The sets are not lexically disjoint: numerals, variable names, and template language overlap, so this experiment tests transfer to new prompt instances rather than a vocabulary-level distribution shift.

6.2 Conjecture Validation Protocol

From the training-set analysis (Section 5), we extract four testable conjectures:

1. **C_scale**: PSI increases monotonically with scale for logic families (binder, theorem, closure)
2. **C2**: Logic families have higher mean PSI than computation families
3. **C_tree**: The 98% training decision tree transfers to held-out tasks (accuracy 80%)
4. **C1**: n_real_win_layers remains the top MI predictor on held-out data

Each conjecture is tested descriptively on the held-out battery across all three models. We report **OBSERVED** when the stated pattern recurs and **WEAKENED** when it does not recur in full; these labels are not significance tests.

6.3 Held-Out Results

The held-out battery contains 54 tasks with no exact prompt overlap with training, across all 6 families. We run PSI on all three models and validate the four conjectures:

Conjecture	Type	Verdict	Key Metric
C_tree	Descriptive threshold transfer	**OBSERVED**	Training 98% → Holdout **86%**
C1	Metric-feature dominance	**OBSERVED**	n_real_win_layers remains top MI predictor
C2	Changed-population family contrast	**OBSERVED**	Logic -0.088 > Compute -0.106
C_scale_binder	Scale monotonicity	WEAKENED	$-0.65 \to -0.35 \to -0.54$
C_scale_theorem	Scale monotonicity	WEAKENED	$+0.13 \to -0.04 \to +0.01$
C_scale_proof	Scale monotonicity	WEAKENED	$-0.04 \to +0.07 \to -0.07$

Overall: three of six pre-specified observations recur, comprising one descriptive threshold observation, one metric-feature observation, and one changed-population family contrast.

The descriptive threshold, metric-feature, and family-separation observations recur on held-out parameterizations. The decision tree trained on 150 training samples achieves 86% accuracy on 162 samples from non-identical prompts built from the same templates; because its primary feature is a direct PSI component and the vocabularies overlap, this does not show that a boundary between mathematical structure and token bias is independently learnable or generalizable.

The per-family scale monotonicity conjectures are weakened. The aggregate held-out contrast is not a like-for-like replication: unlike the 50-task training battery, the 54-task held-out battery includes logical_constraint among the logic-labelled families. Its small +0.0188 difference is therefore reported descriptively, without a significance claim, and remains subject to the same surface-form and control limitations.

6.4 Interpretation

The validation supports the following hierarchy of claims, ordered by evidence strength:

1. **Descriptive:** A threshold rule using a direct PSI component classifies PSI > 0 with >85% accuracy on held-out parameterizations of the same templates.
2. **Descriptive:** The number of correct-winning layers (n_real_win_layers) has the highest measured MI with PSI > 0 because it enters the metric directly.
3. **Descriptive:** Under the expanded held-out logic grouping, its unweighted family mean is 0.0188 higher than the computation grouping.
4. **Limited:** The specific scale-trend is non-monotone and task-dependent; larger task batteries, semantics-breaking tests, paraphrases, and reported ablations are needed for stronger per-family claims.

6.5 Cross-Architecture Validation: Pythia

To test whether the PSI contrast is a GPT-2-family artifact, we repeated the PSI analysis on four GPT-NeoX/Pythia models (Biderman et al., 2023): Pythia-70M, Pythia-160M, Pythia-410M, and Pythia-1B. Because larger Pythia models make pure Python forward passes prohibitively slow, this replication uses a Rust GPT-NeoX tracer that loads safetensors once and emits per-layer first-token option logits in batch mode. For first-token option ranking, replacing full log-probabilities with option logits preserves rankings and margins because the layer-wise log-softmax denominator is common to all options.

The cross-architecture result is a preliminary replication under a first-token proxy:

Model	Mean PSI	Positive PSI	Logic PSI	Compute PSI	Logic−Compute Gap
Pythia-70M	−0.1100	34%	−0.1139	−0.1328	+0.0189
Pythia-160M	−0.1771	28%	−0.1409	−0.2336	+0.0927
Pythia-410M	−0.2154	20%	−0.0224	−0.2520	+0.2295
Pythia-1B	−0.1000	32%	+0.0525	−0.2303	+0.2828

The global mean PSI remains negative: for Pythia, as for GPT-2, at least one reported control matches or exceeds the real-task fraction on average. Every Pythia model has a positive Logic−Compute gap, but the differing first-token measurement protocol means this is only a preliminary cross-architecture comparison. Within the tested Pythia models, the gap is positive and increases across these four scale points; at Pythia-1B, the average logic-family PSI becomes positive while computation-family PSI remains strongly negative.

This provides a preliminary first-token-proxy replication of the GPT-2-family pattern; it does not establish that transformers acquire relational/logical mathematical structure earlier than arithmetic/computational structure.

6.6 Cross-Architecture Hypothesis

Within each measurement protocol, the tested family groups show different PSI profiles. The GPT-2-family gap is positive but non-monotone after 117M; the Pythia first-token-proxy gap is positive and increases across its four tested scale points. We do not pool these series because their scoring protocols differ.

We use **family-dependent acquisition** only as a hypothesis for this descriptive pattern. The present evidence does not establish separate transition curves, grokking, or freedom from surface-form confounds.

7. The Theorem_Transitive Exception

7.1 Option-Scoring Evidence

In GPT-2-medium, the theorem_transitive_01 prompt group ("If A implies B and B implies C, then A implies") shows:

- Internal win rate: 100% across all 5 paraphrases
- Final output accuracy: 100%
- Paraphrase Jaccard: 0.83 (well above the 0.5 threshold)

No other evaluated prompt group in the benchmark achieves this combination.

7.2 Representation Probe (Orthogonal Validation)

To complement option scoring, we designed an auxiliary probe: hidden-state cosine similarity across paraphrases of the same mathematical task, measured at each layer. This probe does not use the unembedding matrix, but it is not independent of lexical or token-level prompt similarity.

Peak-layer paraphrase consistency (maximum over recorded block outputs of the mean pairwise cosine similarity across 5 paraphrases):

Prompt group	Family	distilgpt2	GPT-2	GPT-2-medium
Addition	Arithmetic	0.961	0.953	0.959
Inverse arithmetic	Arithmetic	0.964	0.952	0.965
Cancellation	Algebra	0.954	0.956	0.964
Term collection	Algebra	0.954	0.946	0.963
Variable binding	Binder	0.968	0.955	0.963
Transitivity	**Theorem**	**0.980**	**0.981**	**0.977**
Contradiction closure	Closure	0.944	0.947	0.970
Reflexivity closure	Closure	0.964	0.937	0.968

The prompt labels abbreviate the exact companion IDs listed in Appendix B. theorem_transitive_01 shows the **highest paraphrase consistency** across all three models. This is an auxiliary geometric signal consistent with the option-scoring finding, but not independent of surface-form confounds.

7.3 Family Clustering

We measure within-family vs between-family cosine similarity at each layer to test whether the model organizes prompts by mathematical structure.

Model	Best Clustering Layer	Best Ratio	Pattern
distilgpt2	layer_3 (of 6)	1.17	Mid-network peak
GPT-2	layer_7 (of 12)	1.18	Mid-network peak
GPT-2-medium	embedding	1.25	Embedding-level only

The observed family-clustering ratios are weak (1.02–1.25). In GPT-2-medium, the maximum ratio occurs at the embedding (1.25), drops to 1.03 at block 0, and remains much smaller through the

block outputs. This pattern is consistent with strong surface-form organization at the embedding; it does not by itself identify an active causal mechanism.

7.4 Interpretation

Option scoring and an auxiliary hidden-state geometry signal both show that the theorem-transitive prompt group is unusually stable in these measurements. Because the geometric signal is consistent with, but not independent of, surface-form confounds, this does not show that the model has learned relational structure beyond surface token patterns.

This is not a claim about "understanding." The narrower observation is that this prompt group has unusually stable option-scoring and paraphrase geometry in the battery. Whether that reflects logical encoding or a robust surface pattern remains open.

8. Discussion

8.1 What XRayBench Shows

The primary contribution is methodological: a reusable, controlled protocol for evaluating claims about internal mathematical representations. The protocol is designed to be skeptical because the probe-level null hypothesis—that the reported controls can match the real-task signal—is difficult to reject.

For the three GPT-2-family checkpoints with complete XRay runs, the null hypothesis is not rejected at the global level. Individual controls provide probe-specific challenges:

- **Matched-token control:** produces the largest gap for GPT-2, although control dominance varies by checkpoint.
- **Random-label control:** often matches or exceeds the real-task internal win rate under reassigned correctness labels.
- **Semantics-breaking control:** leaves substantial internal win rates under the implemented perturbations, without isolating semantic content conclusively.

8.2 The PSI Scaling Dichotomy: Implications

The PSI results reveal that asking "does mathematical skill emerge?" is the wrong question. The right question is "which mathematical skills emerge, and at what rate?"

The three logic-requiring families that show higher PSI at some tested scales share a property: they require tracking relationships between entities rather than computing numerical transformations. Binder tracking requires knowing which variable currently holds which value. Theorem patterns require understanding logical implication chains. Proof closure requires recognizing when a proof state matches a closure condition.

In contrast, arithmetic and algebra require numerical computation: adding, multiplying, simplifying. Their mean PSI does not exceed the two PSI controls in the reported measurements, including GPT-2-medium, GPT-2-large, and Pythia-1B.

This pattern is consistent with, but does not establish, different behavior for relational and computational prompts across the tested decoder-only transformer families. Surface-form confounds and the incomplete ablation evidence remain open alternatives.

8.3 What Remains Open

Scale and architecture

1. **Scale threshold.** At what parameter count does global PSI cross zero? Even Pythia-1B remains negative globally, although its logic-family PSI becomes positive.
2. **Additional architectures.** Qwen, Llama, and instruction-tuned models have different training distributions and may show different emergence profiles.

Training and causality

3. **Instruction tuning.** RLHF may reorganize internal representations; the current analysis is limited to base models.
4. **Causal direction.** PSI describes option-scoring differences after two controls; it does not establish genuine structure or causation. Targeted, reported ablation studies could test which model components contribute to the observed Logic−Compute split.

8.4 Limitations

The benchmark operates on completion-format models via direct weight-level forward passes. GPT-2-family runs use pure-numpy forward passes; the Pythia replication uses a Rust GPT-NeoX implementation for tractability. This limits the current implementation to models with publicly available weights in safetensors format and excludes instruction-tuned models. The 50-task battery is designed for interpretability depth rather than statistical breadth; a larger battery would improve confidence intervals on the XRay Score.

The representation probe measures hidden-state similarity, not classification accuracy. Prior work shows both sides of the interpretive risk: hidden-state probes can decode arithmetic errors (Sun et al., 2025), while high linear-probe accuracy can arise from task-format differences rather than the purported reasoning mode (Sahoo et al., 2026). As a first probing-classifier check here, a linear classifier trained on hidden states reaches 86.7%, 86.7%, and 93.3% best-layer test accuracy for distilgpt2, GPT-2, and GPT-2-medium, respectively. With only 50 tasks and no independent template split, this result is partly driven by surface prompt format: the receipts place the first fully separable compute-family layer at layer_0 in all three models, with substantial but incomplete embedding-level separability (0.5952). This probe is therefore an auxiliary, confound-sensitive warning, not an independent novelty claim or a replacement for PSI controls.

8.5 What We Do Not Claim

We do not claim that these results generalize to instruction-tuned models, which may develop different internal representations through RLHF. We do not claim the XRay Score is the only valid metric for internal mathematical skill — it is one operationalization of control-based evidence. PSI removes only the random-label and matched-token option-scoring controls; semantics-breaking, paraphrase, and reported ablation evidence remain necessary to exclude additional confounds. We do not claim that the theorem_transitive exception represents "reasoning" in any cognitive sense — only that its internal representation is measurably more stable than other mathematical families in our benchmark.

8.6 Implications for the Field

We suggest that interpretability research on mathematical reasoning adopt a "controlled-first" methodology: before claiming that a model "knows" something internally, demonstrate that the signal survives matched-token, semantics-breaking, and paraphrase controls. The XRayBench protocol provides a concrete implementation of this principle.

The PSI analysis extends the theorem_transitive observation from one prompt group to a descriptive pattern in the tested option-scoring measurements: the author-defined family groups do not show the same PSI profiles. This does not establish that logical and relational structure is encoded at smaller scales than arithmetic computation. Interpretability studies should report task-family composition and additional controls rather than infer emergence from aggregate scores alone.

For mathematical training curriculum design, the results motivate measuring relational prompts separately from arithmetic transformation rather than collapsing both into a single "math" score. They do not establish comparative training efficiency.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Cursor and its integrated AI models (Claude and GPT families) in order to assist with manuscript drafting, structural editing, and language polishing. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. nostalgebraist (2020). interpreting GPT: the logit lens. *LessWrong*. Primary source
2. Belrose, N. et al. (2023). Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*. https://arxiv.org/abs/2303.08112
3. Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1), 207–219. DOI: 10.1162/coli_a_00422
4. Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems 35*. DOI: 10.52202/068431-1262
5. Saxton, D., Grefenstette, E., Hill, F., & Kohli, P. (2019). Analysing Mathematical Reasoning Abilities of Neural Models. *International Conference on Learning Representations*. https://arxiv.org/abs/1904.01557
6. Lewkowycz, A. et al. (2022). Solving Quantitative Reasoning Problems with Language Models. *arXiv preprint arXiv:2206.14858*. DOI: 10.52202/068431-0278
7. Stolfo, A., Belinkov, Y., & Sachan, M. (2023). A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing. DOI: 10.18653/v1/2023.emnlp-main.435

8. Radford, A. et al. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.
9. Biderman, S. et al. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of Machine Learning Research*, 202, 2397–2430. https://proceedings.mlr.press/v202/biderman23a.html
10. Hou, Y. et al. (2023). Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4902–4919. DOI: 10.18653/v1/2023.emnlp-main.299
11. Razeghi, Y., Logan IV, R. L., Gardner, M., & Singh, S. (2022). Impact of Pretraining Term Frequencies on Few-Shot Reasoning. *arXiv preprint arXiv:2202.07206*. DOI: 10.18653/v1/2022.findings-emnlp.59
12. Sun, Y., Stolfo, A., & Sachan, M. (2025). Probing for Arithmetic Errors in Language Models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8111–8128. DOI: 10.18653/v1/2025.emnlp-main.411
13. Sahoo, S., Jain, V., Chadha, A., & Chaudhary, D. (2026). Linear Probes Detect Task Format, Not Reasoning Mode in Language Model Hidden States. *Proceedings of the 6th Workshop on Trustworthy NLP*, 227–239. DOI: 10.18653/v1/2026.trustnlp-main.12

Appendix A: Implementation Details

GPT-2-family experiments use pure-numpy forward passes without deep learning framework dependencies (Radford et al., 2019). Pythia experiments use a Rust GPT-NeoX tracer for batch per-layer option scoring. The evaluated checkpoints are distilgpt2 , gpt2, gpt2−medium, gpt2−large, EleutherAI/pythia−70m, EleutherAI/pythia−160m, EleutherAI/pythia−410m, and EleutherAI/pythia−1b. Model weights are loaded from Hugging Face safetensors files. The forward_with_intermediates method returns the residual-stream hidden state at the last token position for each recorded position.

Multi-token GPT-2 scoring computes per-token log-probabilities for each option by running the model on prefix + option tokens, extracting log-softmax at each layer via the unembedding matrix (with layer normalization applied before projection). Scores are normalized by option token length to avoid length bias, then further normalized by subtracting scores under a content-free prompt. The Pythia replication uses first-token option logits; for first-token option ranking the omitted log-softmax denominator is shared by all options at a fixed layer, so rankings and margins are preserved.

The representation probe computes cosine similarity between hidden-state vectors across paraphrases. No training is required; the probe is purely geometric.

A hash-bound reproducibility archive accompanies the Zenodo record. The public project mirror is maintained at https://github.com/tnedr/thalens-open.

Appendix B: Full Experimental Data

B.1 Per-Model Phase Summary

See companion files:

- math_skill_full_experiment_distilgpt2.md
- math_skill_full_experiment_gpt2.md
- math_skill_full_experiment_gpt2_medium.md

B.2 Representation Probe Data

See companion files:

- representation_probe_distilgpt2.md
- representation_probe_gpt2.md
- representation_probe_gpt2_medium.md