Latent Depth-Routing Spectroscopy in Standard Transformers: Oracle Evidence on Frozen Gemma-2-2b

Sohail Mohammad
Independent Researcher

Preprint, 2026

Abstract

Standard transformers process inputs through a fixed sequence of layers, but the degree to which different layers contribute to each prediction varies with the input. We introduce oracle-alpha, an optimization-based method that recovers input-dependent effective depth mixtures from the frozen residual stream of `google/gemma-2-2b` without modifying model weights. On a stratified confirm surface of 1024 prompts spanning four task strata, oracle routing improves next-token loss over uniform weighting by +1.6299 nats (95% CI $[1.5865, 1.6758]$; 1024/1024 prompts positive), and held-out predicted routing generalizes positively (+0.8292 nats; 958/1024 positive; $R^2 = 0.2392$). Softmax-constrained competitive routing outperforms matched unconstrained gating by +0.4888 nats (128/128 prompt-level wins) and all tested top-k variants (128/128 wins per k), achieving this advantage with fewer effective sources (34.07 vs 45.59 under unconstrained gating). Routing structure is non-random and task-conditioned: factual recall exhibits strong raw-source clustering (silhouette 0.4709 vs 0.1401 random, 32/32 resamples), with within-family route modes that differentiate by prompt presentation format as well as semantic content, though factual recall also has the lowest held-out positive rate among strata. Extension lanes provide bounded and mixed evidence: tool-breakage shows positive disruption against fixed-alpha controls (+1.0898) but negative results on primary dynamic donor controls, safety mediator activity remains refusal-collapsed on the current quartet surface (preventing broader alignment claims), and the OIH external anchor is positive but calibration-sensitive (+0.0478 dynamic minus calibrated static). We frame these results as frozen-model effective depth-mixture analysis (association, not causal validation) and do not claim trained-router equivalence.

1 Introduction

Transformer language models process every input through the same fixed stack of attention and MLP layers. Yet not all layers contribute equally to every prediction. Early layers may handle syntactic parsing while later layers retrieve factual associations or plan multi-step reasoning. If this heterogeneous contribution pattern is input-dependent, the residual stream of a standard transformer should contain a latent depth-routing signal that varies with the input.

We ask a specific empirical question: can this latent signal be recovered from a frozen model's representations, and if so, what structure does it exhibit?

To answer this, we introduce oracle-alpha, an optimization procedure that finds per-sequence softmax-constrained routing weights over the fixed residual sources of a frozen transformer. These weights specify an effective depth mixture: how much each layer's output should contribute to reconstruction of the model's original predictions. Oracle-alpha operates entirely post-hoc on cached representations and does not modify model weights.

This framing matters for interpretation. A positive oracle-alpha signal establishes that useful input-dependent routing structure is recoverable from standard transformer representations. It does not establish what a co-adapted trained routing system (such as Attention Residuals [3]) would learn, because trained routing involves a feedback loop between routing decisions and learned representations that frozen analysis cannot capture.

Contributions. We provide:

1. A confirm-tranche evaluation showing strong oracle routing signal on frozen Gemma-2-2b across 1024 prompts, with held-out predictiveness (Section 4.1).
2. Evidence that softmax-constrained competitive routing is empirically superior to unconstrained and top-k alternatives, with an effective-source analysis characterizing the mechanism (Section 4.2).
3. Structured, task-conditioned routing patterns with strongest support in factual recall, including within-family route-mode analysis revealing prompt-frame conditioning (Section 4.3).
4. Bounded extension-lane evidence for tool-breakage, safety alignment, and external anchor generalization, with explicit mixed-lane caveats (Section 5).

2 Related work

Depth-adaptive computation. Several lines of work explore adaptive computation depth in transformers. Early exits allow models to skip later layers for easy inputs [18, 17]. LayerSkip trains models with layer dropout to enable early exit at inference [5]. ShortGPT finds that pruning entire layers from pretrained models often preserves performance, suggesting heterogeneous layer importance [10]. The Curse of Depth provides theoretical grounding for why deep layers become increasingly redundant [20]. Our work differs from these approaches by recovering per-input soft routing weights over all layers simultaneously, rather than making binary skip-or-keep decisions.

Mixture-of-Depths. Raposo et al. [16] train models to dynamically allocate computation by routing tokens past entire transformer blocks. Our work examines routing in frozen standard models rather than training new routing-capable architectures.

Residual ensembles and depth-weighted architectures. Residual networks behave as ensembles of relatively shallow networks [21], suggesting that not all paths through the network are equally important. DenseFormer enhances information flow by learning depth-weighted averages across all preceding layers [15]. The Attention Residuals architecture [3] replaces the standard fixed residual stream with learned, input-dependent routing weights that form a softmax-constrained mixture over all preceding layer outputs. Our oracle-alpha method is motivated by Attention Residuals but operates post-hoc on frozen standard models. We do not claim that oracle-alpha recovers what trained Attention Residuals routing would learn after co-adaptation.

Mixture-of-Experts routing. Sparse Mixture-of-Experts models route tokens to different experts within each layer [19, 6, 23]. Our work examines routing across depth (which layer outputs to weight) rather than across width (which expert to select within a layer). The competition insight from our regime comparison (Section 4.2) is related to the load-balancing and expert-utilization literature in MoE.

Representation analysis tools. The logit lens [14] and tuned lens [2] project intermediate representations to vocabulary space to study how predictions form across layers. We use tuned-lens KL divergence as the primary metric for the tool-breakage lane (Section 5.1). Factual knowledge localization methods [11, 12] identify which layers store specific facts. Our factual recall routing patterns (Section 4.3) complement this work by showing that the model routes different factual queries through measurably different depth mixtures. Residual stream circuit analysis [4] provides a mathematical framework for understanding how information flows through transformer layers.

Activation steering and safety features. Representation engineering [24] and activation steering modify internal representations to control model behavior. Refusal in language models has been shown to be mediated by a single direction [1], and harmfulness and refusal are encoded separately [22]. Safety-critical layers have been identified in aligned models [9]. We use refusal direction localization at specific layers as a lens into safety-related routing structure (Section 5.2), without claiming that routing differences imply practical behavioral control.

3 Methods

3.1 Model and prompt surface

We use `google/gemma-2-2b` [7] (2 billion parameters, 26 transformer layers, 53 residual sources per prompt) as the primary frozen model spine. For the safety lane, we use `google/gemma-2-2b-it` (the instruction-tuned variant) to study refusal-related features.

The prompt surface is `registry_v5`, a stratified collection of prompts across four task strata: factual recall (knowledge retrieval questions), reasoning and math (multi-step inference), code and procedural (programming tasks), and general text (open-ended generation). The registry is split into a pilot tranche ($N = 256$) for method selection and hyperparameter tuning, and a confirm tranche ($N = 1024$) for claim-bearing evaluation. All method choices are locked on the pilot tranche before evaluating on confirm.

3.2 Oracle-alpha optimization

For each input sequence, we extract and cache all 53 residual-source outputs (the embedding output, plus attention and MLP outputs at each of the 26 layers). We then optimize a per-sequence routing vector $\mathbf{z} \in \mathbb{R}^{53}$, producing routing weights $\boldsymbol{\alpha} = \text{softmax}(\mathbf{z})$. The routed output is the weighted sum of cached source outputs, passed through the model's original final normalization layer.

The optimization target is the cross-entropy loss of the routed output against the model's original next-token predictions, evaluated under the model's own final normalization. We use Adam optimization for 20 steps with learning rate 0.1 and seed 11.

Reconstruction constraints. Uniform routing ($\alpha_i = 1/53$ for all sources) must reproduce the model's original logits after final normalization to numerical precision. Per-source decomposition uses the shared final normalization factor from the full routed mixture, not per-source normalization.

3.3 Held-out predictiveness

To test whether oracle routing captures generalizable structure (not just in-sample fitting), we train a ridge regression predictor on the pilot tranche. The predictor maps input features (position-thirds

mean-pooled hidden states from layer 4, concatenated with mean-pooled token embeddings) to oracle routing logit vectors. We evaluate predictiveness on the held-out confirm tranche using R^2, mean Jensen–Shannon divergence to oracle $\boldsymbol{\alpha}$, and the fraction of prompts with positive improvement over uniform.

The regularization coefficient is selected from a grid $(10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100)$ using pilot-tranche cross-validation.

3.4 Routing regimes

We compare three routing regime families with matched initialization ($\mathbf{z} = \mathbf{0}$ for all regimes):

- **Softmax-constrained:** $\boldsymbol{\alpha} = \text{softmax}(\mathbf{z})$, producing a probability simplex over sources.
- **Unconstrained:** $\alpha_i = \sigma(z_i)$, allowing each source weight to vary independently without a sum-to-one constraint.
- **Top-k:** $\boldsymbol{\alpha} = \text{softmax}(\text{top}_k(\mathbf{z}))$, retaining only the k largest components before applying softmax. We test $k \in \{2, 4, 8, 13, 26\}$. Because the arg-top$_k$ operation has zero gradients almost everywhere, we use a straight-through gradient estimator.

3.5 Null models

We evaluate oracle routing against four preregistered null baselines:

1. **Uniform:** $\alpha_i = 1/53$ for all sources.
2. **Random Dirichlet:** $\boldsymbol{\alpha}$ sampled from $\text{Dir}(1, \ldots, 1)$.
3. **Magnitude proportional:** α_i proportional to the ℓ_2 norm of source i.
4. **Last layer only:** $\boldsymbol{\alpha}$ concentrated on the final layer output.

3.6 Extension-lane methods

Tool-breakage (Section 5.1). We measure how oracle routing disrupts the model's internal representations using tuned-lens [2] KL divergence. For each prompt in the factual recall subset, we compare the tuned-lens output distribution under original vs oracle-routed representations. We employ four control arms: fixed-alpha (pilot mean), prompt-permuted alpha, within-family permuted alpha, and cross-family permuted alpha. Dynamic donors use seeded-derangement pairing (seed 13) to remove prompt-order artifacts.

Safety alignment (Section 5.2). We use `gemma-2-2b-it` with a quartet design: each prompt group contains refusal, harmful_context, benign, and safe_reply roles. We first localize refusal and harmfulness directions via contrast pairs, then validate behavioral separation, and finally examine mediator-conditioned routing structure by partitioning prompts based on refusal direction activation.

OIH external anchor (Section 5.3). We evaluate oracle-alpha on an external MIB-compatible [13] multiple-choice QA surface (`resattn_oih_mcqa_v1`; $N = 110$ pilot, 50 confirm). We compare dynamic predicted routing against two static baselines: a ShortGPT-style [10] pruned static policy and a pilot-mean-alpha static policy. The calibrated static comparator is the better-performing of the two static policies.

Table 1: Oracle vs null baselines on the confirm tranche ($N = 1024$). *Decision use: the oracle ceiling of +1.63 nats over uniform establishes the recoverable routing signal; the last-layer-only failure (+17.8 oracle advantage) motivates depth-mixture analysis over single-layer approaches.*

Baseline	Mean loss (nats)	Oracle advantage (nats)
Uniform	4.7341	+1.6299
Random Dirichlet	9.4581	+6.3539
Magnitude proportional	6.9846	+3.8804
Last layer only	20.9150	+17.8108
Oracle	**3.1042**	—

Table 2: Oracle signal by stratum on the confirm tranche ($N = 256$ per stratum). *Decision use: factual recall shows the largest oracle gap but lowest predicted positive rate, while reasoning/math shows the smallest gap but highest predicted rate, indicating different routing learnability across task types.*

Stratum	Oracle impr.	Predicted impr.	Predicted positive	Mean JS
Factual recall	+2.2488	+1.6297	239/256 (93.4%)	0.0580
General text	+1.9501	+0.6645	234/256 (91.4%)	0.1374
Code/procedural	+1.4595	+0.5376	238/256 (93.0%)	0.1181
Reasoning/math	+0.8613	+0.4848	247/256 (96.5%)	0.0803

4 Results

4.1 Oracle signal and generalization

On the confirm tranche ($N = 1024$), oracle-alpha routing improves mean next-token loss over uniform weighting by +1.6299 nats (95% bootstrap CI $[1.5865, 1.6758]$). All 1024 prompts show positive improvement. The oracle improves over every preregistered null baseline (Table 1).

Held-out predicted routing (ridge regression trained on the pilot tranche) also yields positive improvement: +0.8292 nats over uniform (95% CI $[0.7805, 0.8758]$; 958/1024 prompts positive). The confirm R^2 between predicted and oracle alpha is 0.2392, and mean JS divergence to oracle alpha is 0.0985. This indicates that the oracle signal reflects recoverable structure, not just in-sample fitting.

Stratum-level breakdown. The oracle signal varies by task stratum (Table 2).

An informative dissociation emerges: factual recall has the largest oracle gap (+2.25) but the lowest predicted positive rate (93.4%), while reasoning/math has the smallest oracle gap (+0.86) but the highest predicted positive rate (96.5%). This is consistent with the interpretation that the oracle finds deeper routing gains on factual recall, but the ridge predictor generalizes more reliably on reasoning/math. The JS-to-oracle values reinforce this pattern: predicted routes are structurally closest to oracle routes for factual recall (JS = 0.058), consistent with factual recall exhibiting the most structured routing.

Table 3: Regime comparison on the confirm surface ($N = 128$). *Decision use: if implementing depth routing, softmax constraint is empirically preferred over unconstrained or top-k; the competition mechanism achieves better loss with fewer effective sources.*

Regime	Impr. (nats)	95% CI	Positive	Eff. sources	Gini
Softmax	**+2.5525**	[2.46, 2.64]	128/128	34.07	0.511
Unconstrained	+2.0637	[1.99, 2.14]	128/128	45.59	0.303
Top-k, k=26	+1.4664	[1.37, 1.57]	128/128	22.95	0.642
Top-k, k=13	+1.0891	[0.99, 1.19]	124/128	13.55	0.773
Top-k, k=8	+0.6765	[0.58, 0.78]	110/128	14.24	0.739
Top-k, k=4	+0.2032	[0.15, 0.27]	47/128	35.01	0.340
Top-k, k=2	+0.0115	[0.00, 0.03]	6/128	50.61	0.045

4.2 Regime comparison and competition mechanism

On the confirm comparison surface ($N = 128$), softmax-constrained routing outperforms every tested alternative on every prompt (Table 3).

Softmax outperforms unconstrained gating by +0.4888 nats on every prompt (128/128 wins), supporting the hypothesis that competition among sources via the simplex constraint is empirically beneficial.

Competition mechanism. The effective-source counts reveal why competition matters. Softmax uses 34.07 effective sources while unconstrained uses 45.59. Softmax achieves a better result with fewer active sources: the simplex constraint forces sources to compete for weight, producing more selective and more useful routing. The unconstrained regime spreads weight more broadly (lower Gini = 0.30) but gains less. This is consistent with the prediction that zero-sum competition concentrates weight on the most informative sources for each input.

Top-k gradient considerations. Top-k routing requires a straight-through gradient estimator because the arg-top$_k$ operation has zero gradients. The monotonic degradation of top-k as k decreases likely reflects optimization difficulty from the straight-through estimator rather than fundamental regime inferiority. The effective-source count for top-k is non-monotonic: k=2 has 50.61 effective sources (nearly all sources receive weight through the softmax applied to the top-2), while k=8 has only 14.24. This non-monotonicity is inconsistent with a clean sparsity interpretation and strongly suggests gradient estimation artifacts dominate the top-k results.

Sparsity profile. Gini coefficients show that the optimal sparsity is moderate. Softmax (Gini = 0.51) outperforms both the more sparse top-k, k=13 (Gini = 0.77) and the less sparse unconstrained (Gini = 0.30). Neither extreme sparsity nor near-uniform spreading is optimal.

4.3 Routing structure and task conditioning

4.3.1 Full-surface structure

Across the full 1024-prompt mixed surface, routing exhibits strong structure at the grouped-source level but not at the raw-source level. The source-type view (embedding vs attention vs MLP) achieves silhouette 0.7558 (vs 0.4681 random; 32/32 resamples oracle > random), and the depth-thirds-by-type view achieves 0.5168 (vs 0.1433 random; 32/32 resamples). Raw-source clustering

Table 4: Stratum-conditioned raw-source clustering ($N = 256$ per stratum). *Decision use: factual recall and reasoning/math show structured routing above the 0.2 silhouette threshold; code and general text do not, limiting extension to those domains.*

Stratum	Oracle sil.	Random sil.	Ratio	Resamples > random	Best k
Factual recall	**0.4709**	0.1401	**3.36x**	32/32	12
Reasoning/math	**0.2456**	0.1401	**1.75x**	32/32	12
Code/procedural	0.1339	0.1401	0.96x	12/32	—
General text	0.1301	0.1401	0.93x	4/32	2

Table 5: Factual recall route modes within the k=12 clustering. *Decision use: route modes differentiate by prompt presentation format within semantic families, suggesting routing captures processing strategy rather than factual content alone.*

Family	N	Modes	Mode sizes	Mean centroid JS
Capitals	64	3	32/16/16	0.1610
Elements	64	3	32/16/16	0.2096
Authors	64	5	18/16/16/13/1	0.1533
Moons	64	2	32/32	0.1770

on the full surface is aggregate-heavy (best k=2, cluster sizes 1023/1, silhouette 0.1664) and does not exhibit meaningful raw-source differentiation.

This indicates a two-level structure: coarse grouped-source organization (the model routes qualitatively differently through attention vs MLP components) is broadly present, while fine-grained raw-source differentiation requires task-specific conditioning.

Mass distribution. Across the full surface, the average routing profile allocates 0.0114 to embeddings, 0.5344 to attention outputs, and 0.4542 to MLP outputs, with mean entropy of 3.5186 bits and 33.83 effective sources.

4.3.2 Stratum-conditioned raw-source structure

When we condition on task stratum, meaningful raw-source clustering emerges in two of four strata (Table 4).

Factual recall shows strong raw-source structure (silhouette 0.47, exceeding the preregistered 0.2 threshold by 2.4x). Reasoning/math shows real but weaker structure (silhouette 0.25, above the 0.2 threshold). Code/procedural and general text do not show significant raw-source structure beyond random controls.

4.3.3 Factual recall route modes

Within the factual recall stratum ($N = 256$), k=12 clustering produces near-perfect family purity: 255/256 prompts cluster with their semantic family, with only one singleton outlier. The four factual families each split into distinct routing modes (Table 5).

Mechanistic differentiation. Route modes differ in their attention-vs-MLP balance, consistent with qualitatively different processing strategies. Capitals cluster 2 (map-style prompts) shows at-

tention mass enrichment ($\Delta = +0.053$), with strongest source weights at layers 17, 2, and 21 (attention outputs). Capitals cluster 4 (quiz-style prompts) shows MLP mass enrichment ($\Delta = +0.021$), with strongest weights at layers 2 and 14 (MLP outputs). Elements cluster 5 (table/abbreviated prompts) shows attention mass enrichment ($\Delta = +0.063$). Author cluster 6 (literature-students prompts) shows a pronounced MLP shift (attention $\Delta = -0.061$, MLP $\Delta = +0.065$).

Prompt-frame conditioning. A key observation is that route modes are not conditioned solely on the semantic content of the query (which capital, which element) but also on the prompt's presentation format. Within the capitals family, map-style prompts (e.g., "The capital shown on this map is...") route differently from quiz-style prompts (e.g., "What is the capital of..."), despite asking about the same domain. This is consistent with the interpretation that the oracle recovers routing structure reflecting how the model processes different presentation formats, not just what factual content is being retrieved.

4.3.4 Reasoning/math as a secondary positive stratum

Reasoning/math achieves silhouette 0.2456, above the preregistered 0.2 threshold, with 100% resampling reliability (32/32). While the structure is weaker than factual recall and we do not present detailed mode analysis for this stratum, the positive finding indicates that depth-routing organization is not unique to knowledge retrieval. Reasoning tasks also show non-trivial input-dependent routing structure.

5 Extension lanes (bounded and mixed evidence)

The following three extension lanes provide additional perspectives on the oracle-alpha signal. Each is reported with explicit caveats reflecting the current evidence state.

5.1 Tool-breakage

The tool-breakage lane tests whether oracle routing measurably disrupts the model's internal representations. If oracle-alpha merely finds weights that trivially reweight outputs without affecting the model's processing, the tuned-lens output distribution should remain unchanged.

Baseline result ($N = 20$ confirm prompts, 10 route modes). Routing increases mean tuned-lens KL divergence by +2.9065 (all 10 route modes positive), and increases target-rank range on 85% of prompts. This shows that oracle routing alters internal representations relative to uniform, but does not yet establish route-specificity (tested below).

Fixed-alpha control. To test whether route-mode-specific structure matters (beyond any-nonuniform routing), we compare oracle routing against a fixed pilot-mean-alpha baseline. Oracle routing increases KL by +1.0898 over fixed-alpha, indicating that the specific routing pattern matters, not just the act of departing from uniform.

Dynamic donor controls. To test whether the disruption is specific to matched (correct) routing rather than any permuted routing, we compare against three dynamic donor arms with seeded-derangement pairing (seed 13; Table 6).

Table 6: Tool-breakage donor-arm decomposition ($N = 20$ prompts, 10 route modes). *Decision use: the +1.09 fixed-alpha advantage supports route-specificity, but mixed dynamic controls limit claims about matched routing superiority over any permuted routing.*

Control arm	Routed minus control (mean tuned KL)
Fixed pilot-mean alpha	+1.0898
Prompt-permuted alpha	−0.0343
Within-family permuted alpha	−0.2084
Cross-family permuted alpha	−0.1344

The dynamic controls are mixed or slightly negative on the pooled primary metric. Removing prompt-order collapse (the seeded derangement) does not convert the result to positive, confirming the mixed result is not merely an ordering artifact.

Family-level decomposition. The mixed result is not uniformly distributed across families. Authors drive the negative sum (contribution −2.73), while elements are positive (+0.69) and capitals are near-zero (+0.002). This heterogeneity is consistent with some route modes genuinely producing matched-routing-specific disruption while others do not.

Interpretation. On this small sample ($N = 20$), oracle routing shows increased disruption relative to fixed-alpha baselines, but the specific advantage of matched dynamic routing over permuted dynamic routing is not supported (negative on pooled metric). The family-level heterogeneity suggests some route modes may produce genuine matched-routing-specific disruption while others do not, but the sample size (5–7 prompts per family after mode splits) limits precision. We present this as bounded negative-to-mixed evidence.

5.2 Safety alignment

The safety lane examines whether routing structure relates to refusal behavior in the instruction-tuned variant (`gemma-2-2b-it`).

Localization. Refusal direction localizes at layer 22 and harmfulness direction at layer 25, with near-orthogonal directions (cosine similarity = −0.0162). This near-orthogonality indicates the model encodes harm detection and refusal execution in separate feature directions at distinct layers, consistent with prior work on refusal [1] and harmfulness [22] direction separation. The localization is stable across three iterations of the prompt surface, suggesting it reflects a robust architectural feature rather than a prompt-construction artifact.

Behavioral validation. On the v3 quartet surface (3 pilot groups, 3 confirm groups, 4 roles each), behavioral separation is perfect: refusal behavior hit rate = 1.0, non-refusal pass rate = 1.0, and confirm pair accuracy = 1.0 for both refusal and harmfulness directions.

Mediator-conditioned routing. Partitioning confirm prompts by refusal direction activation (threshold = 122.08) yields 3 active prompts (all refusal-labeled) and 9 inactive prompts (harmful_context: 3, benign: 3, safe_reply: 3). The mediator-active partition collapses to refusal-only: no role diversity expansion is observed.

Table 7: OIH static comparator analysis ($N = 50$ confirm prompts). *Decision use: the pruned-static catastrophic failure (−3.59 nats) motivates input-dependent routing; the modest dynamic-over-calibrated margin (+0.05 nats) tempers claims about dynamic advantage.*

Static policy	Mean improvement (nats)	Positive prompts
Pruned static (ShortGPT-style)	−3.5916	0/50 (0%)
Pilot-mean-alpha static	+0.1813	49/50 (98%)

Interpretation. Clean localization and behavioral validation show that refusal-related routing structure exists and is measurable. However, the mediator collapse to refusal-only means we cannot claim that routing differences distinguish between harmful contexts and safe responses. Broader mediator-expansion claims are not supported on the current surface.

5.3 Oracle-informed heads external anchor

The OIH lane tests whether oracle-alpha generalizes beyond the primary loss surface to an external multiple-choice QA task (`resattn_oih_mcqa_v1`; $N = 110$ pilot, 50 confirm).

Dynamic routing signal. Predicted improvement over uniform: +0.2290 nats. Oracle improvement over uniform: +0.5709 nats. Confirm R^2: 0.0906.

Static comparator analysis. Two static baselines bracket the interpretation (Table 7).

The pruned-static baseline catastrophically fails on the MCQA surface (0/50 positive), showing that naive static depth-pruning does not merely fail to help but actively destroys performance. This motivates input-dependent routing: the value of depth varies by input, and any routing strategy must account for this variation.

The calibrated static comparator (the better-performing pilot-mean-alpha) achieves +0.1813, producing a dynamic-over-calibrated-static margin of +0.0478 nats (approximately 8% of the oracle improvement ceiling of +0.5709).

Interpretation. Dynamic oracle routing is positive on the external MCQA surface and exceeds the calibrated static comparator, but the margin is modest and calibration-sensitive. The choice of static comparator matters substantially: dynamic minus pruned-static = +3.82 nats, but dynamic minus pilot-mean = +0.05 nats, an 80x difference. This underscores the importance of calibrated comparison in extension-lane analyses.

6 Negative results and blocked lanes

We report three lanes that were preregistered but did not reach claim-bearing status. Including these prevents selective reporting.

6.1 Figure 8 proxy alignment (blocked)

To test whether frozen-model routing reproduces the qualitative Figure 8 patterns documented in the Attention Residuals paper [3] (diagonal dominance, embedding persistence, layer-type specialization), we trained a small Block AttnRes proxy (8 blocks, $d_{\text{model}} \in \{96, 160\}$) on WikiText-2 and WikiText-103.

Table 8: Router distillation results across target geometries. *Decision use: the best $R^2 = 0.41$ falls below the preregistered 0.5 gate; token-level supervision improves over sequence-level by 27% relative, informing future router design.*

Target geometry	Best R^2	Mean JS
Block-compressed	0.1662	0.1205
Block-split (sequence MSE)	0.3236	0.0837
Block-split (token MSE)	0.4099	0.0792
Block-split hybrid (late-third)	0.2555	0.1071

Across nine configurations varying tokenization, model width, training horizon, and regularization, a consistent pattern emerged: the proxy wins at early training (200 steps) but loses to the matched baseline at convergence (1500+ steps). More importantly, the entropy ordering is inverted in every configuration (pre-attention entropy < pre-MLP entropy, opposite the expected pre-attention > pre-MLP for layer-type specialization). A regularization sweep over dropout and weight decay does not rescue this inversion.

Deep embedding persistence exists but is weak (0.10 to 0.20 across experiments), providing partial but insufficient evidence for the full Figure 8 pattern surface. Strong trained-routing alignment claims are blocked.

6.2 Router distillation (blocked)

We trained 2-layer MLP routers to predict oracle alpha from hidden states ($\mathbf{h}_4[t]$) across three target geometries (Table 8).

The best result ($R^2 = 0.4099$) falls below the preregistered readiness gate of $R^2 > 0.5$. Two methodological findings are nonetheless informative: token-level supervision improves R^2 by +0.086 (27% relative) over sequence-level supervision on the same target, and block-split targets produce 2.5x higher R^2 than block-compressed targets, suggesting that the block-level decomposition captures meaningful routing structure that compression obscures. Router distillation readiness is not claimed.

6.3 Training dynamics (out of scope)

No claim-bearing training-dynamics artifact exists for this manuscript cycle. This lane is declared out of scope.

7 Discussion

Summary of findings. Oracle-alpha analysis on frozen `gemma-2-2b` reveals a strong, recoverable, input-dependent effective depth mixture. The signal is positive on every confirm-tranche prompt (1024/1024) and generalizes to held-out prediction (958/1024 positive, $R^2 = 0.24$). Competitive softmax routing is empirically superior to unconstrained and top-k alternatives, achieving better loss with fewer effective sources. Routing structure is task-conditioned, with strongest support in factual recall (silhouette 0.47, 12 clusters with 99.6% family purity) and secondary support in reasoning/math (silhouette 0.25). Within factual families, route modes differentiate by both semantic content and prompt presentation format, revealing routing structure that captures how the model processes different input formats.

Extension lanes provide bounded additional evidence. Tool-breakage shows that oracle routing disrupts internal representations beyond fixed-alpha baselines, but matched-vs-permuted dynamic controls remain mixed. Safety localization is clean and stable across three surface iterations, but mediator-conditioned routing remains refusal-collapsed. The OIH external anchor shows positive dynamic routing on MCQA, exceeding the calibrated static comparator by a modest margin.

What oracle-alpha is and is not. Oracle-alpha characterizes the routing signal recoverable from fixed representations. It establishes an upper bound on how much input-dependent depth routing can improve a frozen model's predictions. It does not establish what a trained depth-routing system would learn, because trained routing involves co-adaptation between routing decisions and learned representations that frozen analysis cannot replicate.

The R^2 of 0.24 between predicted and oracle alpha means the ridge predictor explains approximately one quarter of the variance in oracle routing. The remaining 76% may reflect nonlinear routing structure that linear prediction cannot capture, irreducible noise in the optimization, or features not included in the predictor's input representation. Future work with nonlinear predictors or richer feature sets may close this gap.

Why competition matters. The effective-source analysis provides a mechanistic account of why softmax outperforms unconstrained gating. Under the softmax constraint, sources compete for a fixed probability budget. This competition forces the optimizer to allocate weight to the most informative sources for each input, producing a moderately sparse routing profile (Gini = 0.51). Under unconstrained gating, each source weight is optimized independently, resulting in broader but less discriminative weight distributions (Gini = 0.30, 45.59 effective sources). The result is that softmax achieves a +0.49 nats advantage with 25% fewer effective sources.

This competition effect parallels findings in the Mixture-of-Experts literature, where load-balancing losses prevent expert collapse by encouraging competition for token assignments [19, 6]. In our setting, the simplex constraint plays an analogous role without requiring explicit regularization.

Factual recall as the most structured stratum. The concentration of raw-source routing structure in factual recall (silhouette 0.47 vs 0.13–0.14 for code and general text) is consistent with the hypothesis that factual retrieval engages a relatively narrow set of processing pathways [11, 8], while more open-ended tasks use a broader, less structured distribution of layer contributions. The near-perfect family purity (255/256) within the 12 clusters shows that different factual domains (capitals, elements, authors, moons) route through measurably different layer combinations.

The prompt-frame conditioning within families adds nuance: the model does not simply route "capital queries" through one pathway and "element queries" through another. Rather, the routing is jointly conditioned on what is being asked and how the question is framed. This is consistent with the observation that different prompt templates engage different processing strategies even for the same factual content.

8 Limitations

1. **Single model.** All core results are from frozen `google/gemma-2-2b` (2B parameters). We do not test other model families, sizes, or architectures. Generalization beyond this specific model is unknown.

2. **Post-hoc analysis.** Oracle-alpha operates on cached representations and does not involve training. The routing weights may reflect optimization artifacts of the 20-step Adam procedure rather than latent model structure. The held-out predictiveness check ($R^2 = 0.24$) partially addresses this but does not rule it out.

3. **Predictiveness gap.** The R^2 of 0.24 means the majority of variance in oracle routing is unexplained by the linear predictor. Stronger predictors may close this gap, or the gap may reflect inherent noise.

4. **Extension-lane sample sizes.** Tool-breakage uses $N = 20$ prompts with 10 route modes, and safety uses 12 confirm prompts. Family-level decompositions are based on 6–8 prompts per family. These small samples limit the precision and generalizability of extension-lane findings.

5. **Top-k gradient estimator.** The straight-through estimator used for top-k routing may inflate the apparent inferiority of low-k regimes by introducing optimization difficulty orthogonal to regime quality.

6. **Mediator threshold sensitivity.** The mediator-conditioned routing partition depends on a specific activation threshold (122.08). We do not report sensitivity of the refusal-collapse result to threshold choice.

7. **No causal validation.** Oracle-alpha establishes association between routing weights and loss improvement, not a causal mechanism. The tool-breakage lane provides partial interventional evidence but does not establish full causal claims.

8. **Prompt surface coverage.** The `registry_v5` surface covers four broad task strata but does not include all possible task types. Routing structure on untested domains is unknown.

9. **Single optimization seed.** Oracle-alpha uses a fixed random seed (seed 11) for all optimizations. We do not report variance across seeds, leaving open whether the recovered routing structure is seed-invariant or optimization-dependent. The held-out predictiveness ($R^2 = 0.24$) provides partial evidence against pure artifact status, but direct seed-variance analysis would strengthen this claim.

9 Conclusion

We show that a frozen standard transformer (`google/gemma-2-2b`) exposes recoverable input-dependent effective depth-mixture structure on this single 2B-parameter model, though generalization to other architectures and scales is untested. Oracle routing improves over uniform on every tested prompt, generalizes to held-out prediction ($R^2 = 0.24$, indicating substantial unexplained variance), and shows task-conditioned organization strongest in factual recall. Competitive softmax routing achieves better loss with fewer effective sources than unconstrained or top-k alternatives, providing empirical support for the value of depth competition. Extension lanes offer bounded and mixed evidence for representation disruption, safety-related routing localization, and external task generalization, each with explicit caveats. These results characterize recoverable routing signal in fixed representations and do not establish trained-router equivalence.

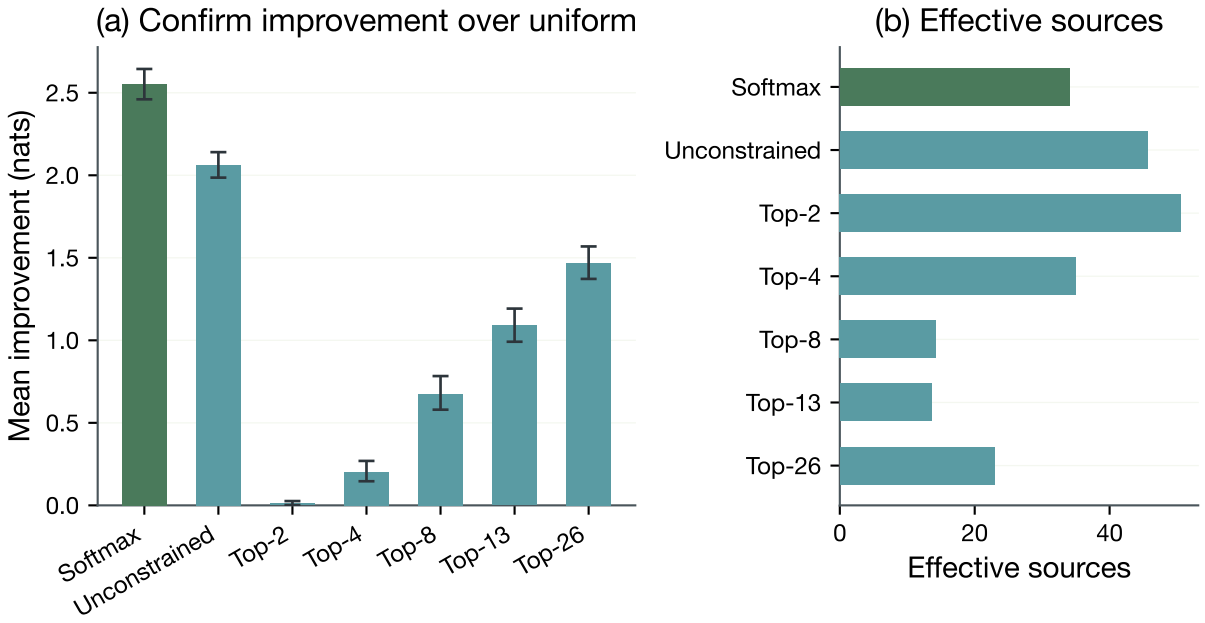

Figure 1: Regime comparison with competition mechanism. Softmax-constrained routing achieves the highest mean improvement (+2.55 nats) with fewer effective sources (34.1) than unconstrained gating (45.6). All pairwise comparisons vs softmax show 128/128 prompt-level wins for softmax. *Decision use: if implementing depth routing, the softmax simplex constraint is empirically preferred; the competition mechanism concentrates weight on informative sources.*

Reproducibility statement

All experiments use fixed random seeds (seed = 11 for oracle optimization, seed = 13 for donor derangements). The prompt surface (`registry_v5`) is frozen before claim-bearing evaluation, with all method choices locked on the pilot tranche ($N = 256$). Result artifacts include SHA256 hashes and are committed to version control. The primary oracle campaign runs on local MPS hardware in approximately 2544 seconds, with checkpoint-based reruns completing in 320 seconds. All null baselines, regime comparisons, and extension-lane controls use the same cached residual sources to ensure comparability. Code and frozen artifacts are available at https://github.com/Sohailm25/latent-depth-routing.

References

[1] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[2] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. Also presented at CoNLL 2023.

[3] Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang,

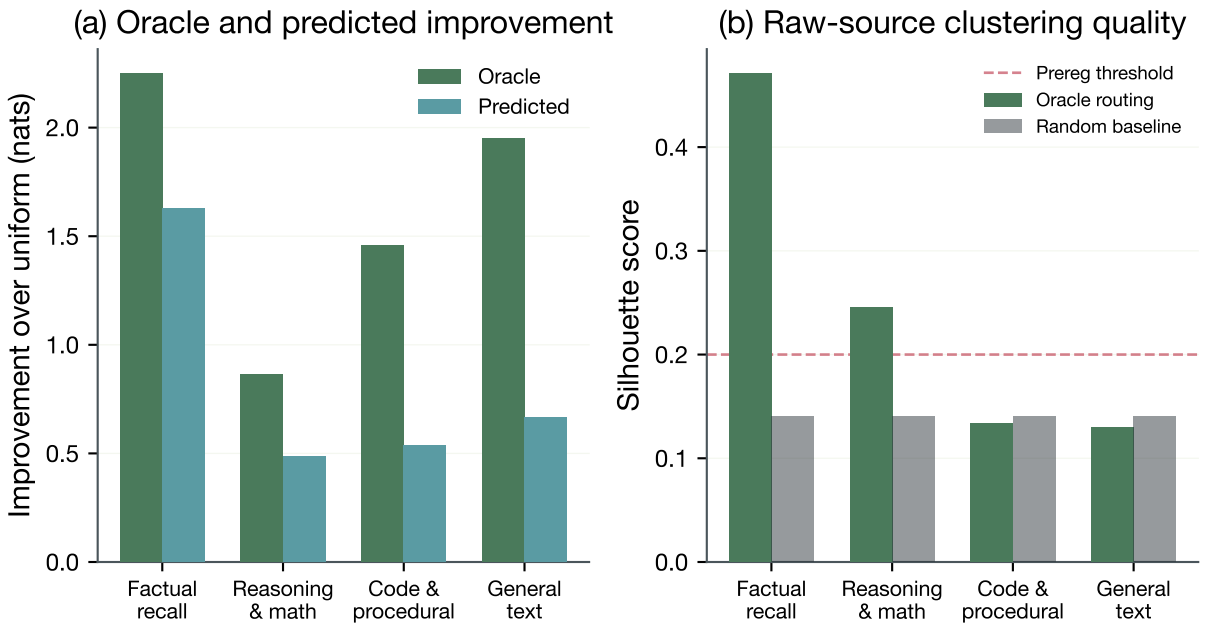

Figure 2: Stratum-conditioned routing structure. (a) Oracle improvement by stratum with predicted improvement overlay. (b) Silhouette scores for raw-source clustering (oracle vs random control). Factual recall shows the largest oracle improvement (+2.25 nats) and strongest raw-source clustering (silhouette 0.47). Reasoning/math shows secondary positive structure (silhouette 0.25). *Decision use: factual recall is the most amenable stratum for depth-routing analysis; code and general text lack raw-source structure.*

Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, et al. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026. Moonshot AI / Kimi Team. Technical report.

[4] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, Anthropic, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[5] Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–40, 2022. arXiv:2101.03961.

[7] Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

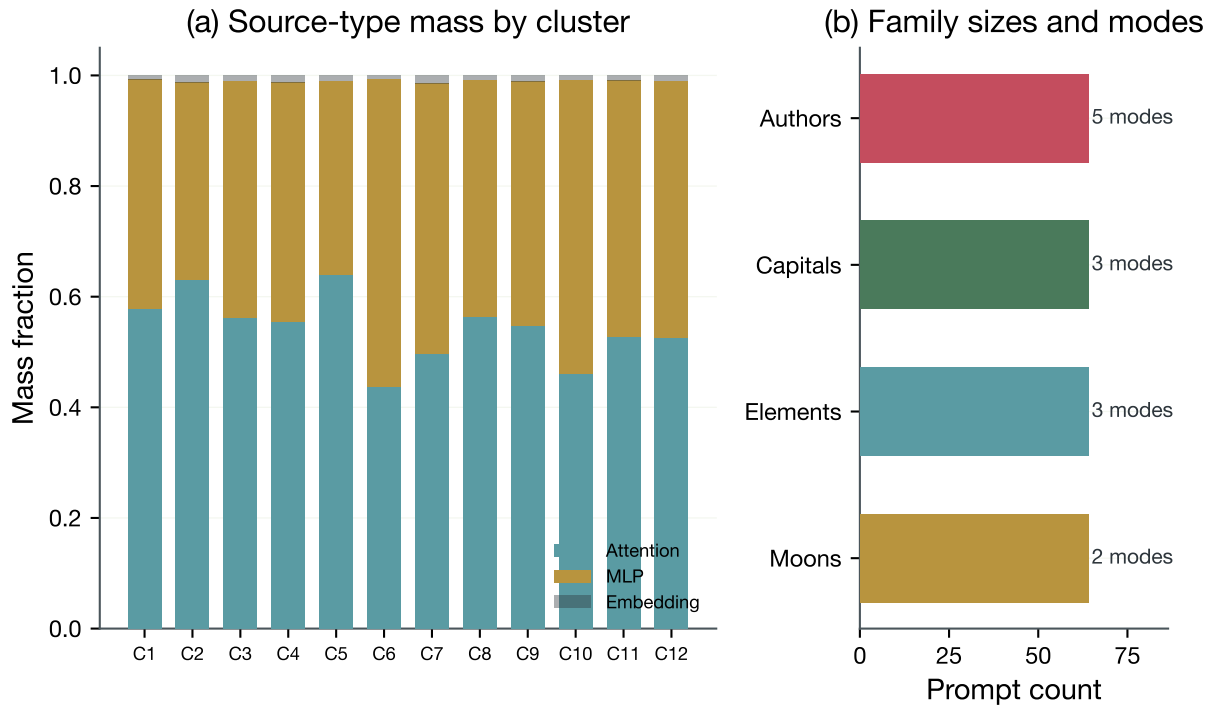

Figure 3: Factual recall route-mode structure. (a) Source-type mass (attention, MLP, embedding) by cluster. (b) Family sizes and mode counts. Within factual recall, k=12 clustering achieves 99.6% family purity (255/256). Route modes differentiate by both semantic family and prompt presentation format. *Decision use: factual queries route through measurably different depth mixtures by both domain and prompt frame, suggesting routing captures processing strategy.*

[8] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12216–12235, 2023. arXiv:2304.14767.

[9] Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.17003.

[10] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2024. arXiv:2403.03853.

[11] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2202.05262.

[12] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.07229.

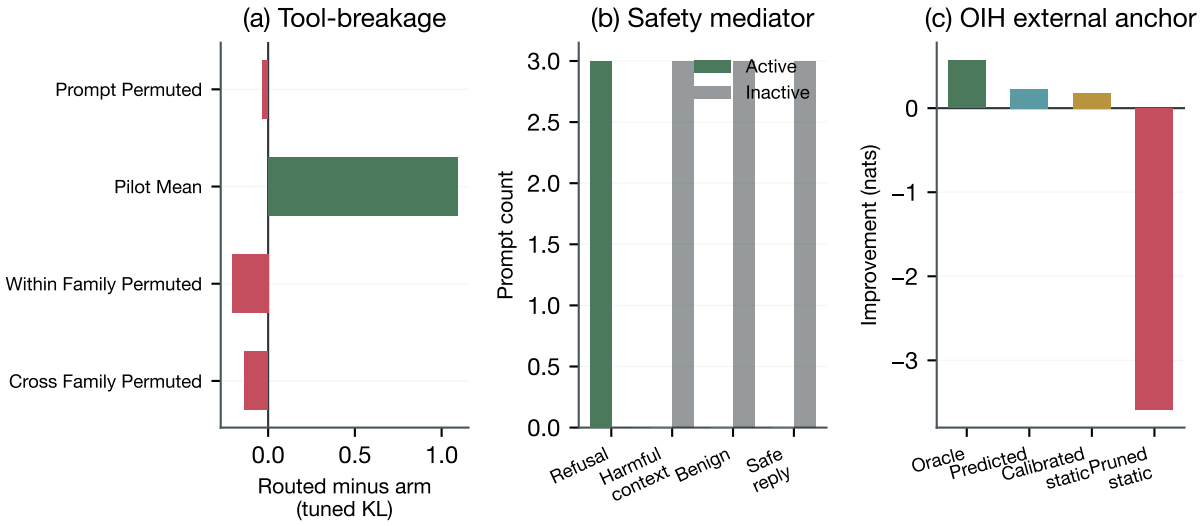

Figure 4: Extension lanes boundary panel. (a) Tool-breakage: oracle routing exceeds fixed-alpha by +1.09 but dynamic donor controls are mixed. (b) Safety mediator: 3 active prompts (all refusal) vs 9 inactive, showing refusal-collapse. (c) OIH external anchor: dynamic routing is positive but the calibrated-static margin is modest (+0.05 nats); pruned static catastrophically fails (−3.59 nats). *Decision use: extension lanes provide bounded evidence with explicit caveats; none supports strong standalone claims.*

[13] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Ivan Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2504.13151.

[14] nostalgebraist. Interpreting GPT: the logit lens. LessWrong blog post, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

[15] Matteo Pagliardini, Amirkeivan Mohtashami, François Fleuret, and Martin Jaggi. DenseFormer: Enhancing information flow in transformers via depth weighted averaging. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. arXiv:2402.02622.

[16] David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2404.02258.

[17] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2207.07061.

[18] Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The right tool for the job: Matching model and instance complexities. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 6640–6651, 2020.

[19] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1701.06538.

[20] Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. arXiv:2502.05795.

[21] Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016. arXiv:1605.06431.

[22] Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. arXiv:2507.11878.

[23] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2202.09368.

[24] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.