UNVEILING THE REASONING PROCESS OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models now solve increasingly complex reasoning tasks, yet the internal computation that transforms token-level inputs into abstract relations remains poorly understood. As further scaling becomes increasingly costly, identifying how reasoning is organized across depth may offer a complementary route toward understanding model intelligence and designing more efficient architectures. Across six base language models and four domains, we first use ΦIDto localize a middle-depth rise in synergistic information. To explain this transition, residual-write geometry reveals opposing and orthogonal updates, suggesting suppression of inherited directions and insertion of new ones. We then test whether these synergy-rich heads matter: rank-guided ablations impair task performance and produce distinct response-length failures. This raises a geometric question—where is the newly composed information represented? Intrinsic-dimension analyses show concurrent expansion of representational support, consistent with the low-dimensional-manifold view that reasoning temporarily recruits additional degrees of freedom. Finally, state transfer shows that composed relations affect downstream computation before direct readout, while Walsh decomposition places higher-order stabilization deeper. Together, these results suggest progressive relational construction rather than passive information transmission, and motivate architectures that differentiate computational roles across depth.

1 INTRODUCTION

Large language models can solve problems that require combining information across many tokens and multiple reasoning steps, yet it remains unclear how these relations are actually formed inside the network. Most layerwise analyses ask either when some information becomes decodable from an intermediate representation or whether intervening on that representation changes the final output (Hewitt & Liang, 2019; Vig et al., 2020; Belrose et al., 2023). These questions are important, but they do not tell us whether the model is simply preserving information, actively reorganizing it, already using a newly formed relation, or merely making that relation easier to read out.

This distinction matters especially for relational reasoning. Knowing the individual pieces of information is not the same as representing their relation. A relation must somehow be constructed from those pieces, become useful to later computation, and eventually be expressed as an answer. These events need not occur at the same layer. We therefore ask a simple question: *how are relations progressively constructed and used across transformer depth?*

Our results reveal a consistent progression. Early layers mainly preserve and propagate existing information. In the middle of the network, residual updates change direction, synergistic information becomes more prominent, and the representation expands across a broader set of directions. This is also where composed relations first begin to affect later computation. Deeper layers then stabilize increasingly higher-order relations and gradually make them directly accessible to the output head. Importantly, a relation can influence the model's computation before it becomes directly readable as an answer. The stages overlap rather than forming sharp boundaries, but together they suggest that depth reflects a gradual construction of relational structure rather than simple information transmission.

We study this process across six dense decoder-only base models from the Qwen3 (Yang et al., 2025), Ministral 3 (Liu et al., 2026), and OLMo 3 (Team OLMo et al., 2026) families, covering mathematics, coding, science, and long-context tasks. We combine information decomposition, residual-write

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.

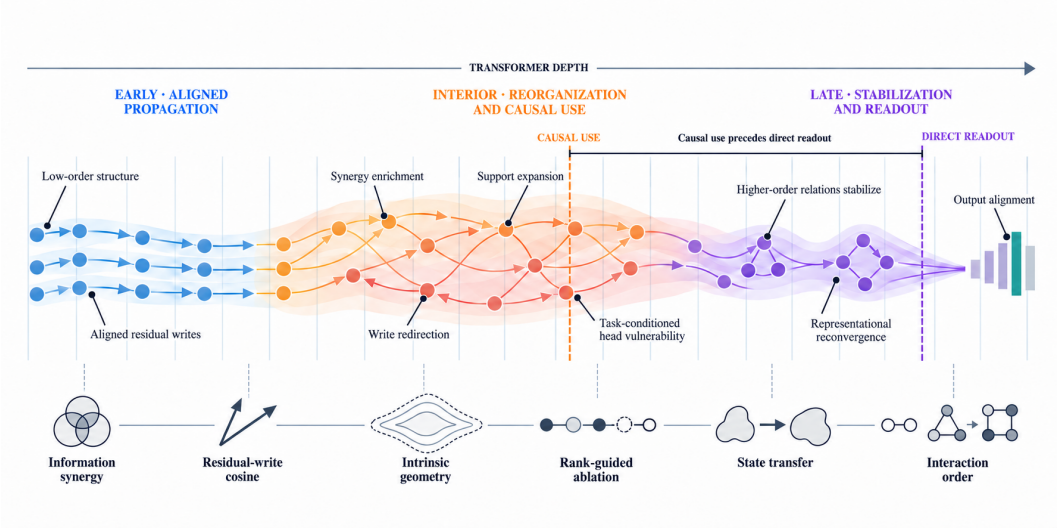


Figure 1: **Progressive relational construction across transformer depth.** Early layers mainly preserve existing information. In the network interior, information is reorganized as residual updates change direction, representational support expands, and composed relations begin to influence later computation.

analysis, representational geometry, head ablation, matched state transfer, and controlled higher-order composition. These analyses examine the same process from complementary perspectives: how information is reorganized, whether that organization matters for behavior, when composed relations become usable, and how relational complexity develops with depth.

Figure 1 summarizes this picture. Rather than locating a single "reasoning layer," we find a broad transition from information preservation, through relational reorganization and causal use, to higher-order structure and direct readout.

The main contributions are:

- **A depth-wise picture of relational reorganization.** Across models and domains, information decomposition, residual-write direction, and representational geometry consistently identify a middle-depth transition in which information becomes more jointly organized and occupies a broader representational space.
- **Evidence that internal use precedes direct readout.** Rank-guided ablations connect this organization to model behavior, while matched state transfer shows that composed relations can affect later computation before the frozen output head can directly recover them.
- **A hierarchy in the formation of relational complexity.** Controlled Fourier–Walsh analysis shows that simpler relations stabilize earlier, whereas higher-order relations requiring more variables to be combined emerge deeper in the network.

2 Related Work

Reading and intervening on intermediate representations. Probing methods ask what information can be recovered from intermediate representations, while approaches such as the tuned lens examine how representations become aligned with the model's output space (Hewitt & Liang, 2019; Belrose et al., 2023). Causal mediation, activation patching, and related interventions instead ask whether changing an internal state changes later behavior (Vig et al., 2020; Meng et al., 2022; Ghandeharioun et al., 2024). Decodability and causal influence, however, need not emerge at the same depth. Our experiments directly compare these two events and show that composed relations can affect downstream computation before they become directly readable.

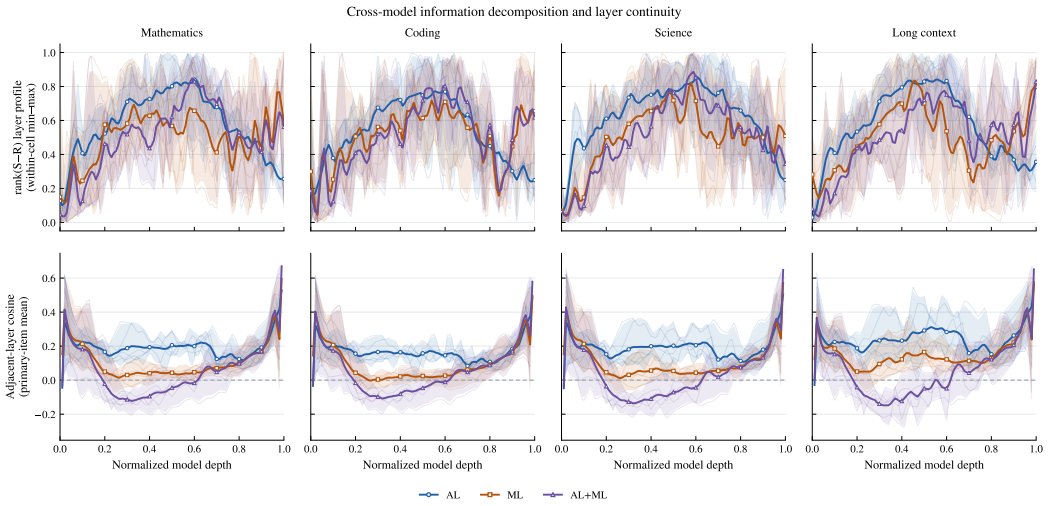


Figure 2: **Middle layers show stronger synergy together with a change in residual-write direction.** Across mathematics, coding, science, and long-context tasks, $\mathrm{rank}(S - R)$ rises toward the network interior while adjacent AL+ML writes become less aligned and often reverse direction. The two changes occur in a similar depth range, suggesting that stronger relational integration emerges together with a substantial rewriting of the residual representation. Curves show the six-model mean, with shaded regions indicating the 2.5th–97.5th percentile range across models.

Head specialization and representational geometry. Previous work has shown that attention heads differ substantially in function and importance, and that many can be removed with limited effects on overall performance (Michel et al., 2019; Voita et al., 2019; Abnar & Zuidema, 2020). Information decomposition provides a complementary view by separating redundant from synergistic information (Williams & Beer, 2010; Barrett, 2015; Mediano et al., 2025), while intrinsic-dimension methods characterize how representations expand and compress during neural computation (Facco et al., 2017; Ansuini et al., 2019). We use these tools together to examine how information organization changes with depth and whether the resulting structure predicts functional sensitivity.

Residual computation and higher-order relations. Transformer blocks repeatedly modify a shared residual stream, making the sequence of residual writes a natural way to study how representations evolve across depth (Vaswani et al., 2017; Elhage et al., 2021). Fourier–Walsh decompositions provide a complementary way to isolate interactions of different orders (Walsh, 1923; O'Donnell, 2014; Tsang et al., 2018), and prior mechanistic work has shown that neural networks can learn structured Fourier-like computations (Nanda et al., 2023). We use controlled relational tasks to ask when interactions of increasing order become established inside pretrained language models.

3 Analytical Framework

We characterize transformer computation through residual writes, information decomposition, and representational geometry.

3.1 Residual writes

Let $\mathbf{h}_{\ell,t}$ denote the residual-stream state entering block ℓ at token position t, and let $\mathbf{a}_{\ell,t}$ and $\mathbf{m}_{\ell,t}$ denote the attention and MLP residual writes, respectively. Following the standard transformer residual formulation (Vaswani et al., 2017; Elhage et al., 2021),

$$\mathbf{h}^{\mathrm{attn}}_{\ell,t} = \mathbf{h}_{\ell,t} + \mathbf{a}_{\ell,t}, \qquad \mathbf{h}_{\ell+1,t} = \mathbf{h}^{\mathrm{attn}}_{\ell,t} + \mathbf{m}_{\ell,t} = \mathbf{h}_{\ell,t} + \mathbf{w}_{\ell,t}, \quad \mathbf{w}_{\ell,t} = \mathbf{a}_{\ell,t} + \mathbf{m}_{\ell,t}. \tag{1}$$

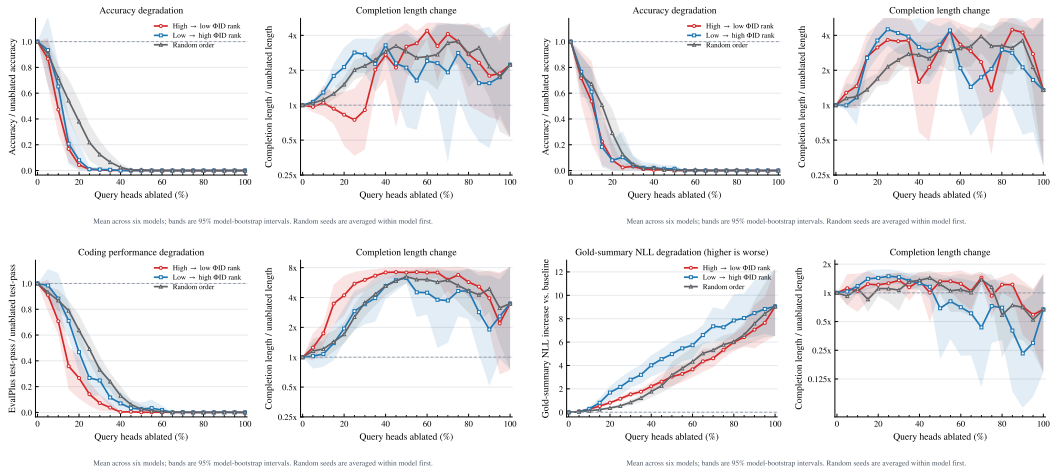


Figure 3: **Rank-guided head ablation reveals task-dependent sensitivity.** The upper row shows GSM8K and ScienceQA; the lower row shows HumanEval+ and GovReport. Each original task panel reports the task endpoint and relative response length on 100 held-out items per model. Curves compare descending and ascending $\mathrm{rank}(S - R)$ with three fixed random orders; lines average six models and ribbons are 95% model-bootstrap intervals.

We call $\mathbf{a}_{\ell,t}$ the *attention-layer write* (AL), $\mathbf{m}_{\ell,t}$ the *MLP-layer write* (ML), and $\mathbf{w}_{\ell,t}$ the *combined write* (AL+ML). At head resolution, $\mathbf{a}_{\ell,h,t}$ denotes the hth query-head value vector immediately before the attention output projection.

3.2 Joint Information Across Token Transitions

Mutual information (MI) measures the statistical dependence between variables, but does not specify how information is distributed across multiple sources. Partial information decomposition (PID) separates the information that two sources X and Y provide about a target Z into redundant, unique, and synergistic components:

$$\begin{aligned} I(X, Y; Z) &= R + U_X + U_Y + S, \\ I(X; Z) &= R + U_X, \qquad I(Y; Z) = R + U_Y. \end{aligned} \tag{2}$$

Here, redundancy R is information about Z available from either source individually, whereas synergy S is information available from the joint observation of X and Y but not from either source alone.

PID describes a static source–target relation. To examine how these information components evolve across model computation, we use integrated information decomposition (ΦID), which extends the same decomposition to a transition between two multivariate states. For two nodes u and v, we consider their scalar trajectories across consecutive response-token positions, $(x_{u,t}, x_{v,t}) \to (x_{u,t+1}, x_{v,t+1})$. ΦID decomposes the mutual information between these joint past and future states according to the PID category of the information in the past and in the future. This produces atoms indexed by transitions between information types. In this work, we use the $s \to s$ atom as synergy and the $r \to r$ atom as redundancy.

More precisely, let $\mathcal{V}$ denote the set of nodes within one model–domain–proxy cell, with $u, v \in \mathcal{V}$ and $u \neq v$. For response item i, applying ΦID to the complete token trajectory of node pair (u, v) gives $S_{uv}^{(i)} = \Phi_{s\to s}^{(i)}(u, v)$ and $R_{uv}^{(i)} = \Phi_{r\to r}^{(i)}(u, v)$, where $S_{uv}^{(i)}$ and $R_{uv}^{(i)}$ are respectively the $s \to s$ and $r \to r$ atoms estimated for that node pair on item i. The explicit construction of $\Phi_{s\to s}^{(i)}(u, v)$ and $\Phi_{r\to r}^{(i)}(u, v)$, including the full ΦID lattice and atom-recovery procedure, is provided in Appendix C.

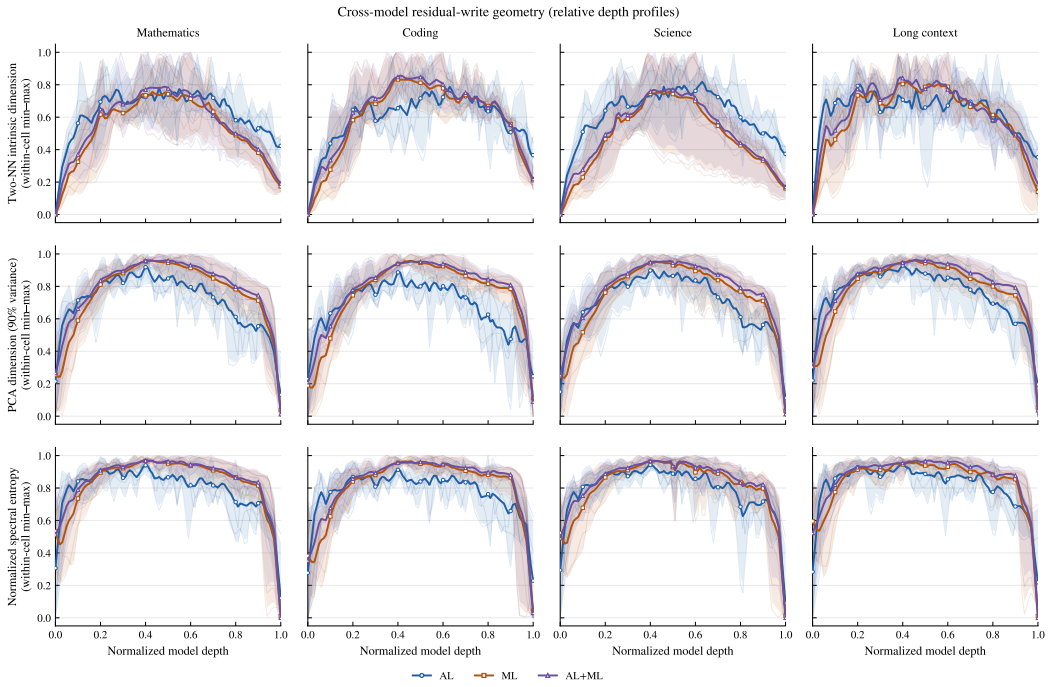


Figure 4: **Residual representations expand in the network interior and contract again toward the output.** Across mathematics, coding, science, and long-context tasks, TwoNN dimension, PCA-90, and spectral entropy rise through the middle layers and decline near the output. The agreement across local-neighbor, variance-support, and spectral measures suggests that the interior representation temporarily occupies a broader set of directions before becoming more concentrated again at later layers. Colors denote AL, ML, and AL+ML; curves show the six-model mean with 2.5th–97.5th percentile bands across models.

We then assign each node v a synergy and redundancy score by averaging first across the N response items and then across all other nodes u:

$$S_v = \frac{1}{|\mathcal{V}|-1} \sum_{\substack{u \in \mathcal{V} \\ u \neq v}} \frac{1}{N} \sum_{i=1}^{N} S_{uv}^{(i)}, \qquad R_v = \frac{1}{|\mathcal{V}|-1} \sum_{\substack{u \in \mathcal{V} \\ u \neq v}} \frac{1}{N} \sum_{i=1}^{N} R_{uv}^{(i)}. \tag{3}$$

Here, N is the number of response items, i indexes an item, and u indexes the partner node of node v. This item-first aggregation gives every response equal weight.

Finally, we compare the two components through $q_v = \mathrm{rank}(S_v - R_v)$, where ranks are computed over nodes within the same model–domain cell. A larger $S_v - R_v$ therefore indicates that node v participates relatively more strongly in generating new informations.

3.3 Representational Geometry

We use TwoNN (Facco et al., 2017) to estimate intrinsic dimension. For each representation x_i, let $r_{i,1}$ and $r_{i,2}$ denote the Euclidean distances to its first and second nearest neighbors, and define $\mu_i = r_{i,2}/r_{i,1}$. Under the locally uniform manifold assumption, the distance ratio follows $F(\mu) = 1 - \mu^{-d}$, which gives $-\log[1 - F(\mu)] = d \log \mu$. We therefore compute the empirical distribution $F(\mu_i)$ over the observed distance ratios and fit a linear relation between $\log \mu_i$ and $-\log[1 - F(\mu_i)]$; the fitted slope gives the intrinsic dimension d. Larger d indicates that the representations occupy a higher-dimensional local manifold.

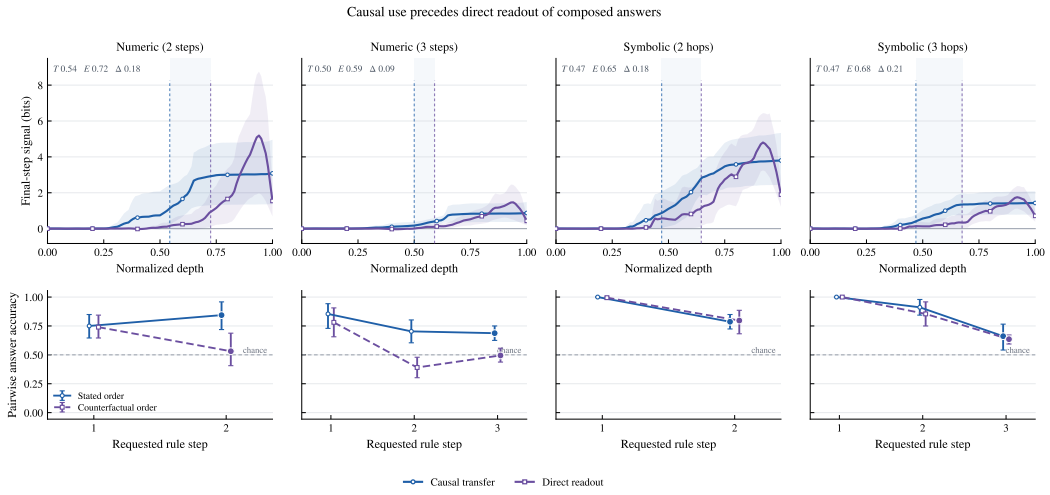


Figure 5: **Composed relations begin to guide computation before they become directly readable.** Across both numeric and symbolic composition tasks, state transfer begins to shift downstream predictions at earlier layers than the same relation becomes recoverable through the frozen output head. This gap is consistent across two- and three-step compositions, suggesting that relational information first becomes functionally useful in an internal form and is only later aligned with the answer space. Lower panels confirm that the models solve both the factual and counterfactual conditions reliably. Ribbons show 95% model-bootstrap intervals, and dashed lines mark the median stable onset of causal use and direct readout.

4 Experimental Design

We first identify a recurring depth-wise reorganization across models and domains, and then test whether this regime is functionally involved in subsequent computation.

4.1 Models, tasks, and depth alignment

The common panel contains Qwen3-8B and Qwen3-14B (Yang et al., 2025), Ministral 3-8B and Ministral 3-14B (Liu et al., 2026), and OLMo 3-7B and OLMo 3-32B (Team OLMo et al., 2026). All are dense base checkpoints spanning 7.30–32.23B parameters and 32–64 blocks; Appendix A gives pinned revisions. Native layers are mapped to normalized depth $d = \ell/(L-1)$.

The structural survey uses Omni-MATH (Gao et al., 2024), BigCodeBench v0.1.4 (Zhuo et al., 2025), OlympiadBench English physics (He et al., 2024), and GovReport (Huang et al., 2021). A deterministic 100-item panel was formed per domain under an 8,192-token budget, and the full panel constitutes the analysis set. Complete solutions or summaries are replayed with teacher forcing, and measurements use response tokens only.

4.2 Structural measurements on naturalistic trajectories

Because ΦID requires scalar series, nodes are write magnitudes: $x^{\mathrm{ML}}_{\ell,t} = \|\mathbf{m}_{\ell,t}\|_2$, $x^{\mathrm{AL+ML}}_{\ell,t} = \|\mathbf{w}_{\ell,t}\|_2$, and $x^{\mathrm{AL}}_{(\ell,h),t} = \|\mathbf{a}_{\ell,h,t}\|_2$. AL is analyzed by query heads; ML and AL+ML have one node per layer.

Directional continuity uses the full write vectors. For $p \in \{\mathrm{AL}, \mathrm{ML}, \mathrm{AL}+\mathrm{ML}\}$,

$$C^{(p)}_{\ell} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{\langle \mathbf{v}^{(p)}_{i,\ell,t}, \mathbf{v}^{(p)}_{i,\ell+1,t}\rangle}{\|\mathbf{v}^{(p)}_{i,\ell,t}\|_2\|\mathbf{v}^{(p)}_{i,\ell+1,t}\|_2}, \tag{4}$$

where N is the number of response items, i indexes an item, T_i is the number of response-token positions in item i, and t indexes those positions. $\mathbf{v}^{(p)}$ denotes $\mathbf{a}$, $\mathbf{m}$, or $\mathbf{w}$ for AL, ML, and AL+ML,

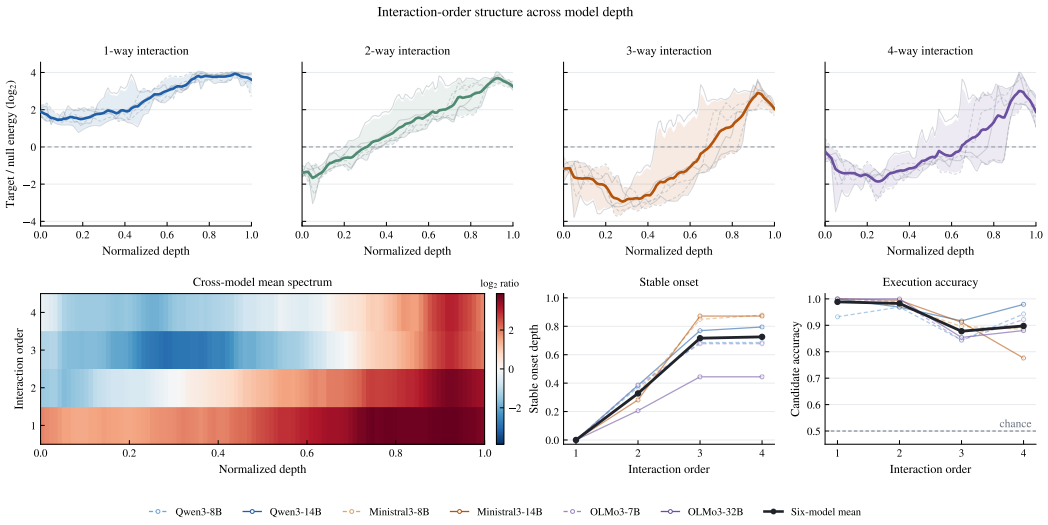


Figure 6: **Higher-order relations stabilize progressively deeper in the network.** Relations involving more variables emerge later across models: first- and second-order interactions stabilize relatively early, whereas third- and fourth-order structure appears only at greater depth. The separation from the shuffled-label baseline shows that this increase reflects the target relation rather than overall activation magnitude, while high candidate accuracy confirms that the later onset is not caused by task failure. Thin curves show individual models, and bold curves summarize the cross-model mean.

respectively. Each item receives equal weight. Positive, near-zero, and negative values indicate aligned, approximately orthogonal, and opposing consecutive writes, respectively.

4.3 Intrinsic geometry

AL, ML, and AL+ML geometry is estimated at every native layer from stratified draws of 50 response positions per item. Intrinsic dimension is computed using the TwoNN estimator introduced in Section 3.3. For the same centered point clouds, let $\lambda_1 \geq \cdots \geq \lambda_D$ denote the covariance eigenvalues. PCA-90 is defined as $k_{0.9} = \min\{k : \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{D} \lambda_j \geq 0.9\}$, the minimum number of principal components explaining 90% of the variance. Normalized covariance-spectrum entropy is $H = -\sum_{j=1}^{D} p_j \log p_j / \log D$, where $p_j = \lambda_j / \sum_{k=1}^{D} \lambda_k$; larger H indicates a more evenly distributed spectrum. Curves are min–max normalized within each model–domain–metric–proxy cell only for Figure 4.

4.4 Held-out tests of head-level functional organization

Following head-pruning analyses (Michel et al., 2019; Voita et al., 2019), we cumulatively ablate attention heads according to descending or ascending $\mathrm{rank}(S - R)$, with random orders as controls. Ranking and evaluation examples are disjoint, allowing us to test whether the identified head organization generalizes beyond the data used to construct the ranking. We evaluate the resulting behavioral changes through task performance and response length, using predictive accuracy for free-generation tasks and gold-summary loss for GovReport.

4.5 Separating causal use from direct readout

We use two controlled composition families. In the numeric setting, a starting digit x is transformed by two or three non-commuting rules $r_1, \ldots, r_h$ ($h \in \{2, 3\}$). The full condition follows

$$x, \quad r_1(x), \quad r_2(r_1(x)), \quad \ldots, \quad r_h(\cdots r_2(r_1(x))), \tag{5}$$

whereas the matched control applies the same rules in reverse order $r_h, \ldots, r_1$. For example, from 2, "double, then add one" gives $2 \to 4 \to 5$, while the reversed order gives $2 \to 3 \to 6$. We retain only cases in which the two trajectories remain single digits and differ at every queried step.

In the symbolic setting, the full prompt contains two disjoint chains,

$$\begin{aligned} s &\to a_1 \to \cdots \to a_h, \\ d &\to b_1 \to \cdots \to b_h, \end{aligned} \tag{6}$$

while the control swaps only the first connections, yielding $s \to b_1$ and $d \to a_1$. Thus, full and control prompts preserve tokenized length and token multisets but imply different targets at every depth. One-hop answers follow an explicit edge, whereas two- and three-hop answers require composing relations not directly stated in the prompt. Each item is probed independently at each composition depth.

At each layer, the final-position block output is passed through the frozen final normalization and LM head. Let $g(\mathbf{z}) = [z(y) - z(y_{\mathrm{cf}})]/\log 2$ denote the logit preference for the full-prompt target y over the counterfactual target y_{cf}. Direct answer evidence and bidirectional state-transfer effects are

$$\begin{aligned} E_\ell &= \tfrac{1}{2}\left[g(\mathbf{z}_\ell^{\mathrm{full}}) - g(\mathbf{z}_\ell^{\mathrm{control}})\right], \\ T_\ell &= \tfrac{1}{2}\left[g(\mathbf{z}_\ell^{c\leftarrow f}) - g(\mathbf{z}^{\mathrm{control}}) + g(\mathbf{z}^{\mathrm{full}}) - g(\mathbf{z}_\ell^{f\leftarrow c})\right]. \end{aligned} \tag{7}$$

Here, E_ℓ measures answer information directly accessible through the frozen LM head. For T_ℓ, the layer-ℓ final-position state is exchanged between the matched full and control prompts and the remaining layers run normally; $c \leftarrow f$ and $f \leftarrow c$ denote the two transfer directions. Thus, $T_\ell > 0$ indicates that the transferred state already affects downstream computation toward the donor target. An earlier onset of T_ℓ than E_ℓ therefore indicates causal use before direct readout.

4.6 Tracing the formation of higher-order relations

We use controlled XOR tasks to trace when relations of different interaction orders emerge. Each task contains four binary variables $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \{0,1\}^4$ and asks for the XOR of a selected subset A, $y_A(\mathbf{x}) = \bigoplus_{j\in A} x_j$. The interaction order is $|A|$, ranging from one to four; higher order therefore means that more variables must be jointly combined to determine the target. For each selected subset, we enumerate all $2^4 = 16$ possible input assignments.

At every layer, we record the final-position block write $\mathbf{w}_\ell(\mathbf{x}) = \mathbf{h}_\ell^{\mathrm{out}}(\mathbf{x}) - \mathbf{h}_\ell^{\mathrm{in}}(\mathbf{x})$. Using the Fourier–Walsh decomposition (Walsh, 1923; O'Donnell, 2014), we isolate the component associated with the selected relation:

$$\begin{aligned} \widehat{\mathbf{w}}_{\ell,A} &= 2^{-4} \sum_{\mathbf{x}\in\{0,1\}^4} (-1)^{\oplus_{j\in A} x_j}\, \mathbf{w}_\ell(\mathbf{x}), \\ e_{\ell,A} &= D^{-1}\, \|\widehat{\mathbf{w}}_{\ell,A}\|_2^2, \end{aligned} \tag{8}$$

where D is the hidden width. Here, $\widehat{\mathbf{w}}_{\ell,A}$ represents the layer-ℓ write associated with the selected XOR relation, and $e_{\ell,A}$ measures how strongly that relation is expressed at this layer.

To verify that the measured energy reflects the target XOR relation, we compare it with a shuffled input–label baseline. A relation is considered formed when the original energy remains higher than this baseline across consecutive layers. Final-layer accuracy is used only to confirm successful task execution.

5 Results

5.1 Middle layers reorganize information into expanded functional representations

Figure 2 shows that $\mathrm{rank}(S - R)$ rises toward middle depth in 24 model–domain cells, indicating stronger synergistic organization in the network interior. In the same region, AL+ML directional continuity develops a pronounced trough in all 24 cells: negative cosine indicates opposing writes, consistent with suppressing or overwriting existing directions, whereas near-zero cosine indicates approximately orthogonal writes that introduce information along new directions.

This reorganization is accompanied by a change in representational geometry. As shown in Figure 4, TwoNN dimension, PCA-90, and spectral entropy all increase toward middle depth and decrease

again near the output. The agreement across these three measures suggests that, while information is being reorganized, the model temporarily recruits a broader set of representational directions before compressing them toward the output.

Figure 3 further links this organization to behavior. In mathematics and coding, ablating high-$\mathrm{rank}(S-R)$ heads first causes the fastest performance collapse relative to both low-$\mathrm{rank}(S-R)$ and random ablation; coding responses simultaneously become several times longer, while mathematical responses show both shortening and later excessive lengthening. ScienceQA shows a similar early accuracy loss, accompanied by strong response-length inflation under rank-guided ablation. GovReport reverses this pattern: removing low-$\mathrm{rank}(S-R)$ heads first produces the largest increase in summary loss and, at higher ablation levels, markedly shorter generations. Thus, the ablations reveal a functional dissociation between the two forms of head organization. Synergy-dominant heads appear to support the integration and transformation of information required by reasoning-intensive tasks, since their removal causes earlier performance collapse and often destabilizes generation length. In contrast, the stronger sensitivity of GovReport to redundancy-dominant heads suggests that preserving and repeatedly carrying task-relevant information is more important for long-context summarization. The accompanying shifts toward abnormally short or prolonged responses further indicate that these heads regulate not only answer quality but also the dynamics of generation.

5.2 Composed relations are used before they are directly readable

A composed relation can influence later computation before it is directly expressed in the model's output space. In the matched transfer experiment, intermediate states already shift the final prediction toward the transferred relation before the frozen LM head can decode that relation itself (Figure 5). This ordering holds across both numeric and symbolic compositions, with a median separation of 0.156 normalized depth. This suggests that relational information first becomes useful in an internal form that later layers can operate on, and only afterward is converted into directly readable answer evidence. The middle layers therefore appear to support active relational computation rather than merely preparing the final output.

5.3 Depth progressively builds higher-order relations

Relations of different complexity do not stabilize at the same depth. First- and second-order relations emerge earlier, whereas third- and fourth-order relations stabilize substantially deeper in the network (Figure 6). Their mean normalized formation depths are 0.000, 0.327, 0.716, and 0.726 for orders one through four, while final task accuracy remains high across orders. This ordering suggests that depth progressively builds relational complexity: simpler dependencies are available early, whereas relations requiring more variables to be combined emerge only after further transformations.

6 Discussion and Limitations

Our results suggest that transformer depth supports a progression from information preservation to relational reorganization, causal use, and later higher-order stabilization. In particular, composed relations begin to influence downstream computation before they become directly readable through the output head, indicating that internal reasoning states need not already resemble the final answer.

The ablation results further suggest a functional distinction between synergy-dominant and redundancy-dominant heads. Reasoning tasks are more sensitive to the former, consistent with a role in integrating and transforming information, whereas long-context summarization is more sensitive to the latter, consistent with preserving task-relevant information over long contexts. Changes in response length also indicate that these disruptions affect generation dynamics, not only final accuracy.

These conclusions remain limited to six dense base models and the tasks studied here. The ΦID and geometric measurements are relative within model–domain cells, whole-state transfer localizes causal use without identifying the underlying circuit, and the higher-order analysis relies on controlled XOR relations. Extending the analysis to post-trained, mixture-of-experts, and more open-ended reasoning settings remains an important direction.

7 CONCLUSION

Our results suggest that transformer reasoning unfolds gradually across depth rather than appearing at a single layer. Early layers mainly preserve existing information, middle layers reorganize and combine it, and deeper layers stabilize more complex relations and make them directly available for prediction. A key finding is that composed relations begin to affect later computation before they can be directly read out from the model. This suggests that reasoning is not simply about storing an answer internally, but about progressively constructing representations that later layers can use and eventually express.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. URL https://arxiv.org/abs/2005.00928.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6109–6119, 2019. URL https://arxiv.org/abs/1905.12784.

Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Physical Review E*, 91(5):052802, 2015. doi: 10.1103/PhysRevE.91.052802.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017. doi: 10.1038/s41598-017-11873-y.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, et al. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024. URL https://arxiv.org/abs/2410.07985.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436, 2021. doi: 10.18653/v1/2021.naacl-main.112. URL https://aclanthology.org/2021.naacl-main.112/.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026. doi: 10.48550/arXiv.2601.08584. URL https://arxiv.org/abs/2601.08584.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Pedro A. M. Mediano, Fernando E. Rosas, Andrea I. Luppi, Robin L. Carhart-Harris, Daniel Bor, Anil K. Seth, and Adam B. Barrett. Toward a unified taxonomy of information dynamics via integrated information decomposition. *Proceedings of the National Academy of Sciences*, 122(39), 2025. doi: 10.1073/pnas.2423297122.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pp. 14014–14024, 2019. URL https://arxiv.org/abs/1905.10650.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. URL https://arxiv.org/abs/2301.05217.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Team OLMo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, et al. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2026. doi: 10.48550/arXiv.2512.13961. URL https://arxiv.org/abs/2512.13961.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByOfBggRZ.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

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

J. L. Walsh. A closed set of normal orthogonal functions. *American Journal of Mathematics*, 45(1): 5–24, 1923. doi: 10.2307/2387224.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. URL https://arxiv.org/abs/1004.2515.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, et al. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YrycTjllL0. Oral presentation.

A Checkpoint Inventory

Table 1: Pinned base-model checkpoints used throughout the study; Q/KV gives query and key–value head counts. Architecture metadata follows the official Qwen3, Ministral 3, and OLMo 3 reports (Yang et al., 2025; Liu et al., 2026; Team OLMo et al., 2026).

Checkpoint	Revision	Params. (B)	Layers	Width	Q/KV heads	Context
Qwen3-8B	`49e3418f`	8.19	36	4,096	32/8	32,768
Qwen3-14B	`0b0bd373`	14.77	40	5,120	40/8	32,768
Ministral 3-8B	`d4883f9b`	8.92	34	4,096	32/8	262,144
Ministral 3-14B	`5b0ceedb`	13.95	40	5,120	32/8	262,144
OLMo 3-7B	`a81bae42`	7.30	32	4,096	32/32	65,536
OLMo 3-32B	`c2b61dae`	32.23	64	5,120	40/8	65,536

B Experiment Inventory

Table 2: Scope and reasoning role of the five primary experiment families and the supplemental layer-skip analysis.

Experiment	Data and unit	Primary measurement	Evidential role
ΦID and cosine	6 models × 4 domains; 100 frozen responses per cell	$\mathrm{rank}(S - R)$ from one-step token dynamics; adjacent full-write cosine	Locate structural reorganization
Head lesions	4 domains; 100 held-out items per model; 21 cumulative doses	Task endpoint and response length	Test functional relevance and task dependence
Layer skips	6 models × 4 domains; 100 frozen responses and 50 fixed response positions per response	Downstream-state relative ℓ_2 change	Trace perturbation propagation across depth
Intrinsic geometry	Same 24 structural cells; 100 items and 10 draws of 50 response positions per item	TwoNN, PCA-90 dimension, normalized spectral entropy	Measure expansion and reconvergence
Matched composition	4 conditions × 32 tasks per model; 320 step probes	Frozen-head readout and bidirectional state transfer	Separate causal use from direct readout
Interaction order	4 orders × 12 frames × 16 assignments per model	Exact target Walsh energy versus 64 permutations	Order relation formation by complexity

C Complete ΦID Construction and Estimation

This appendix makes explicit the construction used in Section 3.2, from the scalar residual-write trajectories to the ΦID atoms and the final node-wise ranking statistic. The purpose is to distinguish

the information-theoretic decomposition itself from the subsequent aggregation used in our cross-model analyses.

C.1 Node trajectories and one-step transitions

For a response item i, let T_i denote the number of response-token positions retained after teacher forcing. As defined in the main text, the scalar trajectory associated with node v is the magnitude of the corresponding residual write,

$$x_{\ell,t}^{\mathrm{ML}} = \|\mathbf{m}_{\ell,t}\|_2, \qquad x_{\ell,t}^{\mathrm{AL+ML}} = \|\mathbf{w}_{\ell,t}\|_2, \qquad x_{(\ell,h),t}^{\mathrm{AL}} = \|\mathbf{a}_{\ell,h,t}\|_2. \tag{9}$$

For a pair of distinct nodes (u, v), we define the two-dimensional past and future states at consecutive response positions as

$$\mathbf{X}_{uv,t}^{(i)} = \begin{bmatrix} x_{u,t}^{(i)} \\ x_{v,t}^{(i)} \end{bmatrix}, \qquad \mathbf{Y}_{uv,t}^{(i)} = \begin{bmatrix} x_{u,t+1}^{(i)} \\ x_{v,t+1}^{(i)} \end{bmatrix}, \qquad t = 1, \ldots, T_i - 1. \tag{10}$$

Thus, ΦID is applied to the one-token transition $\mathbf{X}_{uv,t}^{(i)} \to \mathbf{Y}_{uv,t}^{(i)}$. Importantly, each response item is treated as a separate trajectory: information quantities are estimated within an item and are only aggregated across items after the atom values have been recovered. This prevents long responses from receiving greater weight solely because they contain more token transitions.

C.2 The ΦID product lattice

For two variables, the ordinary PID redundancy lattice contains four information modes (Williams & Beer, 2010):

$$\mathcal{A} = \{r, u_1, u_2, s\}, \tag{11}$$

where, in antichain notation,

$$r = \{\{1\}, \{2\}\}, \qquad u_1 = \{\{1\}\}, \qquad u_2 = \{\{2\}\}, \qquad s = \{\{1, 2\}\}. \tag{12}$$

Here, r denotes information redundantly available from either variable, u_1 and u_2 denote information uniquely available from one variable, and s denotes synergistic information available only from the two variables jointly.

ΦID extends this decomposition from a single source–target relation to a transition between a multivariate past and a multivariate future (Mediano et al., 2025). The corresponding lattice is the Cartesian product

$$\mathcal{A}_\Phi = \mathcal{A} \times \mathcal{A}, \tag{13}$$

so that every atom is indexed by an ordered pair $\alpha \to \beta$, with $\alpha, \beta \in \mathcal{A}$. The first index specifies how the information is distributed across the two variables in the past, and the second specifies how the same information is distributed across the two variables in the future. The bipartite system therefore has $4 \times 4 = 16$ possible information transitions:

past\future	r	u_1	u_2	s
r	$r \to r$	$r \to u_1$	$r \to u_2$	$r \to s$
u_1	$u_1 \to r$	$u_1 \to u_1$	$u_1 \to u_2$	$u_1 \to s$
u_2	$u_2 \to r$	$u_2 \to u_1$	$u_2 \to u_2$	$u_2 \to s$
s	$s \to r$	$s \to u_1$	$s \to u_2$	$s \to s$

. (14)

The partial order on this product lattice is inherited componentwise from the ordinary PID lattice:

$$\alpha \to \beta \preceq \alpha' \to \beta' \quad \Longleftrightarrow \quad \alpha \preceq \alpha' \text{ and } \beta \preceq \beta'. \tag{15}$$

The total time-delayed mutual information between the joint past and joint future is exactly decomposed into the 16 atoms,

$$I(\mathbf{X}_{uv,t}; \mathbf{Y}_{uv,t}) = \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} I_\partial^{\alpha \to \beta}. \tag{16}$$

C.3 DOUBLE REDUNDANCY AND ATOM RECOVERY

The lattice specifies how information atoms are related but does not by itself determine their numerical values. Following the minimum-mutual-information (MMI) construction for Gaussian systems and its multitarget extension to ΦID (Barrett, 2015; Mediano et al., 2025), we define the cumulative double-redundancy quantity associated with a lattice node $\alpha \to \beta$ as

$$I_{\cap}^{\alpha\to\beta} = \min_{\substack{A\in\alpha\\ B\in\beta}} I\big(\mathbf{X}_{uv,t}^{A}; \mathbf{Y}_{uv,t}^{B}\big)\,, \tag{17}$$

where $\mathbf{X}^A$ and $\mathbf{Y}^B$ denote the subvectors selected by the index sets A and B, respectively. When an antichain contains only the joint set $\{1, 2\}$, the corresponding argument is the complete two-variable state.

The cumulative quantity at a lattice node contains the atom at that node together with all atoms below it:

$$I_{\cap}^{\alpha\to\beta} = \sum_{\alpha'\to\beta' \preceq \alpha\to\beta} I_{\partial}^{\alpha'\to\beta'}. \tag{18}$$

Consequently, the individual ΦID atoms are recovered uniquely by Möbius inversion over the product lattice:

$$I_{\partial}^{\alpha\to\beta} = I_{\cap}^{\alpha\to\beta} - \sum_{\alpha'\to\beta' \prec \alpha\to\beta} I_{\partial}^{\alpha'\to\beta'}. \tag{19}$$

Equivalently,

$$I_{\partial}^{\alpha\to\beta} = \sum_{\substack{\alpha'\preceq\alpha\\ \beta'\preceq\beta}} \mu(\alpha', \alpha)\, \mu(\beta', \beta)\, I_{\cap}^{\alpha'\to\beta'}, \tag{20}$$

where μ is the Möbius function associated with the ordinary four-node PID lattice. Equations equation 17–equation 20 completely determine all 16 atoms from the required mutual-information quantities.

The two atoms used in the main analysis have especially direct interpretations. Because $r \to r$ is the bottom element of the product lattice,

$$\begin{aligned} R_{uv}^{(i)} &= I_{\partial}^{r\to r} \\ &= \min\Big\{I(x_{u,t}; x_{u,t+1}), I(x_{u,t}; x_{v,t+1}), I(x_{v,t}; x_{u,t+1}), I(x_{v,t}; x_{v,t+1})\Big\}. \end{aligned} \tag{21}$$

It measures information that is redundantly available across the two nodes in both the past and the future.

Conversely, $s \to s$ is the top element of the product lattice. Its cumulative information is the full time-delayed mutual information,

$$I_{\cap}^{s\to s} = I((x_{u,t}, x_{v,t}); (x_{u,t+1}, x_{v,t+1}))\,, \tag{22}$$

and the atom retained in our analysis is obtained after subtracting all lower-order contributions:

$$S_{uv}^{(i)} = I_{\partial}^{s\to s} = I_{\cap}^{s\to s} - \sum_{\substack{(\alpha,\beta)\in\mathcal{A}\times\mathcal{A}\\ (\alpha,\beta)\neq(s,s)}} I_{\partial}^{\alpha\to\beta}. \tag{23}$$

Thus $S_{uv}^{(i)}$ isolates information that requires the joint state of the two nodes in the past and remains jointly represented in the future, after redundant and unique lower-lattice contributions have been removed.

C.4 GAUSSIAN MUTUAL-INFORMATION ESTIMATOR

The node trajectories are continuous scalar quantities. We therefore estimate the mutual-information terms entering Equation equation 17 using the Gaussian estimator (Barrett, 2015). For jointly Gaussian random vectors $\mathbf{A}$ and $\mathbf{B}$ with covariance matrices $\Sigma_{\mathbf{A}}$, $\Sigma_{\mathbf{B}}$, and joint covariance $\Sigma_{(\mathbf{A},\mathbf{B})}$,

$$I_{\mathrm{G}}(\mathbf{A}; \mathbf{B}) = \frac{1}{2} \log_2 \frac{\det \Sigma_{\mathbf{A}}\, \det \Sigma_{\mathbf{B}}}{\det \Sigma_{(\mathbf{A},\mathbf{B})}}. \tag{24}$$

All covariance matrices for response item i are estimated from its $T_i - 1$ consecutive past–future samples in Equation equation 10. The use of base-2 logarithms expresses the resulting information quantities in bits.

For each node pair, the required scalar-to-scalar, scalar-to-joint, joint-to-scalar, and joint-to-joint mutual informations are first obtained from Equation equation 24. These values determine the 16 cumulative double-redundancy terms through Equation equation 17, after which the 16 atoms are recovered by the Möbius inversion in Equation equation 19. The decomposition is therefore performed independently for every response item and every node pair rather than on a trajectory formed by concatenating different responses.

We use MMI as an operational decomposition rather than assuming that it is the unique possible definition of redundancy. As with PID, ΦID requires a choice of redundancy function, and alternative choices can induce different numerical atom values (Mediano et al., 2025). Our conclusions therefore rely primarily on the reproducible depth-wise organization of the resulting scores, their within-cell ordering, and their independent behavioral and geometric validation, rather than on interpreting an absolute atom magnitude as a model-independent quantity.

C.5 From pairwise atoms to node scores

For every response item i and node pair (u, v), the procedure above produces

$$S_{uv}^{(i)} = I_{\partial}^{s \to s}(u, v), \qquad R_{uv}^{(i)} = I_{\partial}^{r \to r}(u, v). \tag{25}$$

We first average each pairwise quantity over response items and only then over the partner nodes:

$$S_v = \frac{1}{|\mathcal{V}| - 1} \sum_{\substack{u \in \mathcal{V} \\ u \neq v}} \frac{1}{N} \sum_{i=1}^{N} S_{uv}^{(i)}, \qquad R_v = \frac{1}{|\mathcal{V}| - 1} \sum_{\substack{u \in \mathcal{V} \\ u \neq v}} \frac{1}{N} \sum_{i=1}^{N} R_{uv}^{(i)}. \tag{26}$$

This ordering of averages gives every response item equal weight, independently of its token length. The statistic used throughout the paper is then

$$q_v = \mathrm{rank}_{v \in \mathcal{V}} \left(S_v - R_v \right). \tag{27}$$

The ranking is computed within the corresponding model–domain–proxy cell. Hence the analysis asks whether a node is more synergy-dominant or redundancy-dominant relative to comparable nodes in the same model and task, rather than comparing raw information magnitudes across architectures with different widths, depths, or activation scales.

C.6 Interpretation and scope

Two distinctions are important for interpreting these quantities. First, $s \to s$ and $r \to r$ are dynamical information atoms, not ordinary single-target PID terms: $s \to s$ denotes information that is jointly available across the node pair in the past and remains jointly available after one token transition, whereas $r \to r$ denotes information that is redundantly available across the pair at both times.

Second, the variables entering the decomposition are residual-write magnitudes rather than the full hidden vectors. The ΦID analysis therefore localizes changes in the organization of write-magnitude dynamics; by itself it does not identify the semantic content of a representation or establish a causal computational circuit. This is why the main paper uses ΦID as a structural localization criterion and evaluates its functional relevance independently through held-out rank-guided head ablations, residual-write geometry, and controlled state-transfer experiments.

D Additional Evidence for Depth-Wise Organization

D.1 Attention-head ΦID structure

The layer-level trend in the main text does not mean that all heads in a middle layer behave alike. The head maps below show substantial variation within the same layer, but high-$\mathrm{rank}(S-R)$ heads appear more often in the network interior across models and domains. The middle-depth peak therefore comes from a greater concentration of synergy-dominant heads, rather than a uniform shift of every head in those layers.

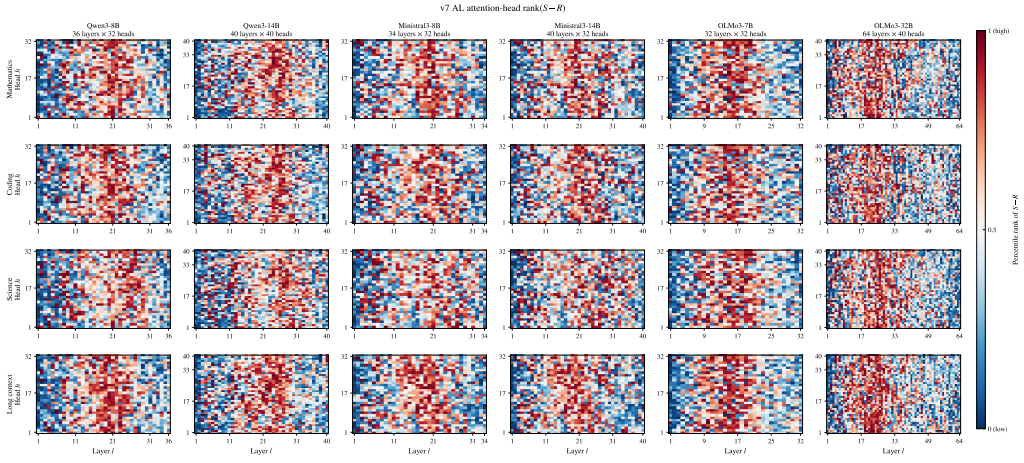


Figure 7: **Head-level** $\mathrm{rank}(S-R)$ **across models and domains.** High-$\mathrm{rank}(S-R)$ heads are more concentrated in middle layers, while strong within-layer variation remains. Rows correspond to mathematics, coding, science, and long context; columns correspond to the six checkpoints. Color shows the within-cell percentile rank of $S-R$.

D.2 WHOLE-LAYER SKIP SENSITIVITY

We also ask how far a perturbation at one depth propagates through the remaining network. For each frozen prompt–response trace, we skip source block ℓ by passing its input directly to the next block and then run the rest of the model normally. The resulting change in later hidden states is measured for every valid source–target pair.

The maps show two recurring patterns. Early blocks often affect a long span of later layers, while some architectures contain additional bands of strong influence in the middle or near the end of the network. This makes the middle-depth transition less like an isolated "critical layer" and more like one part of a distributed computation whose downstream reach depends on both depth and architecture.

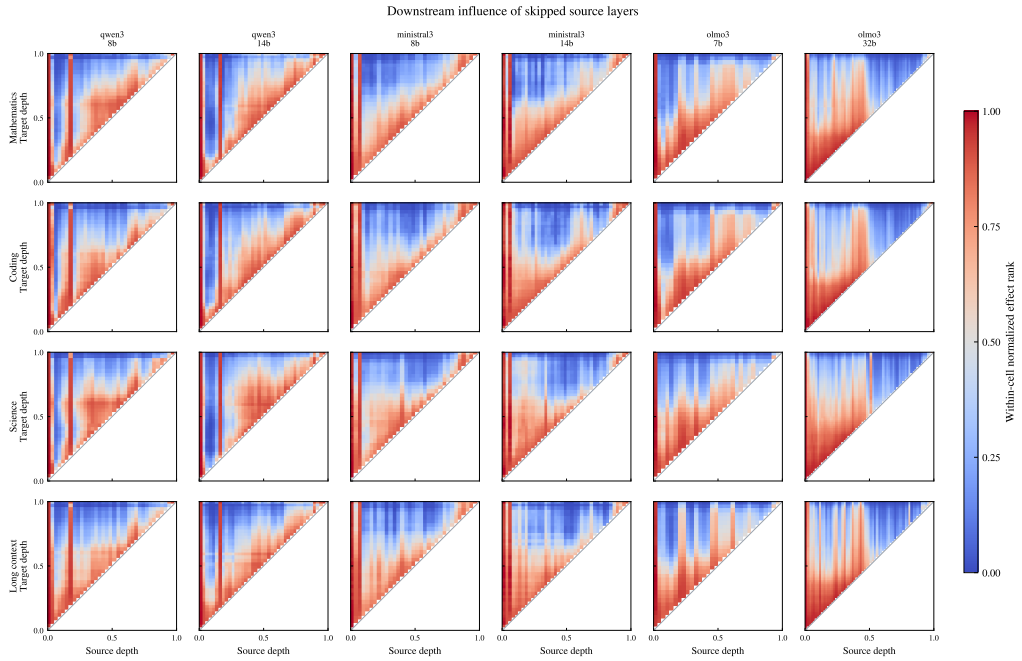


Figure 8: **How a skipped layer changes later hidden states.** Each triangular map shows the normalized rank of the downstream hidden-state change for a skipped source layer and a later target layer. Early layers typically have broad downstream influence, with additional middle- or late-depth bands in some models. Rows correspond to the four task domains and columns to the six checkpoints.

E PER-TASK HEAD-ABLATION TRAJECTORIES

The main text reports the cross-model ablation trend; the figures below show the individual model trajectories. They are useful because a similar drop in task score can come from very different generation failures. On GSM8K, HumanEval+, and ScienceQA, high-to-low $\mathrm{rank}(S - R)$ ablation usually harms performance earlier than random removal, but the response can either collapse to unusually short outputs or grow toward the generation limit. GovReport shows the opposite ranking: low-to-high $\mathrm{rank}(S - R)$ ablation gives the largest gold-summary NLL increase in all six models. The ranking therefore reflects task-dependent functional organization rather than a single global notion of head importance.

E.1 MATHEMATICAL REASONING

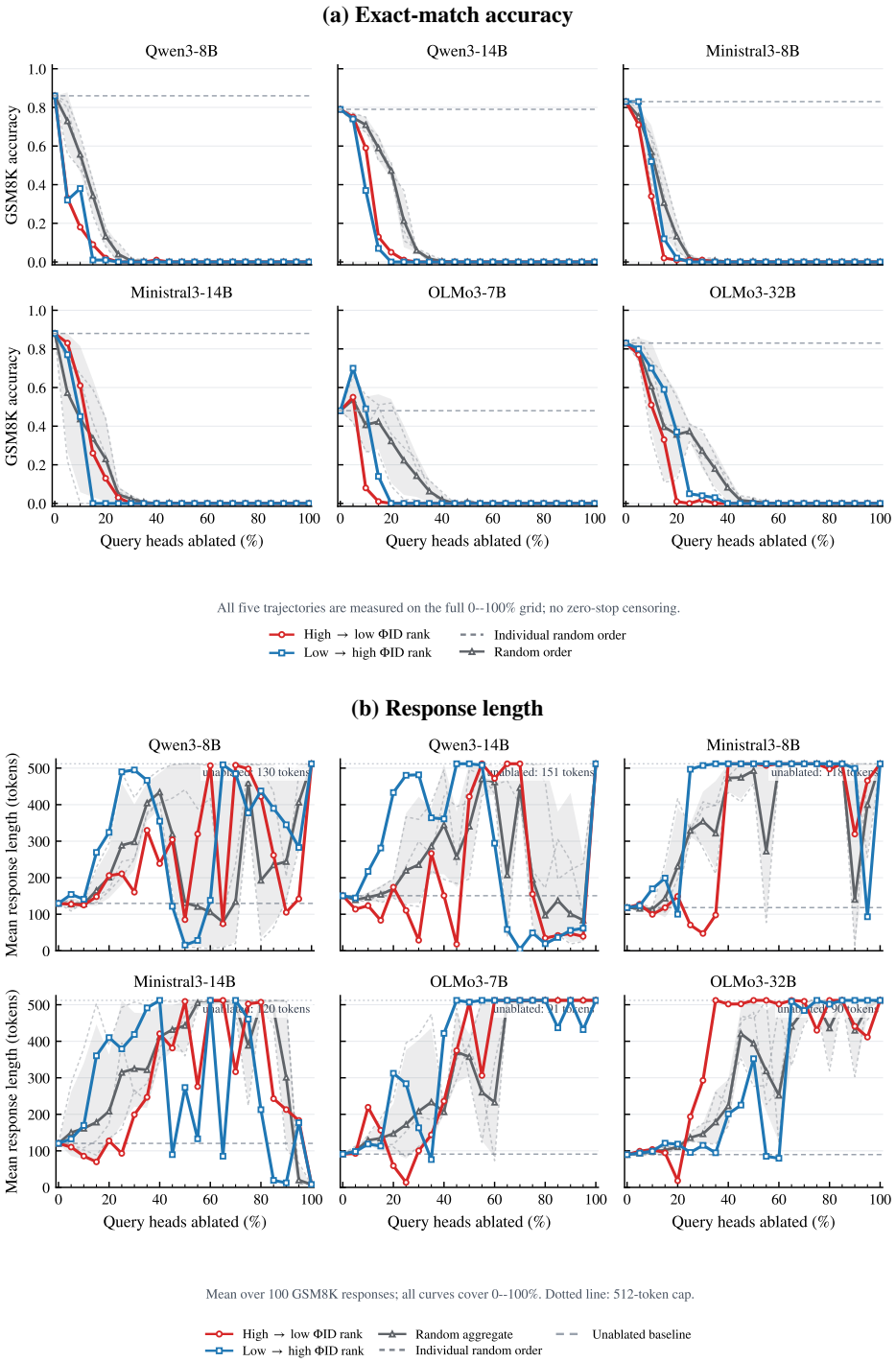


Figure 9: **GSM8K ablation trajectories.** High-to-low $\mathrm{rank}(S - R)$ removal generally causes earlier accuracy degradation than random ablation. Response length reveals that the same accuracy loss can be accompanied by either premature shortening or later excessive generation. Gray curves show the random controls.

E.2 SCIENCE QUESTION ANSWERING

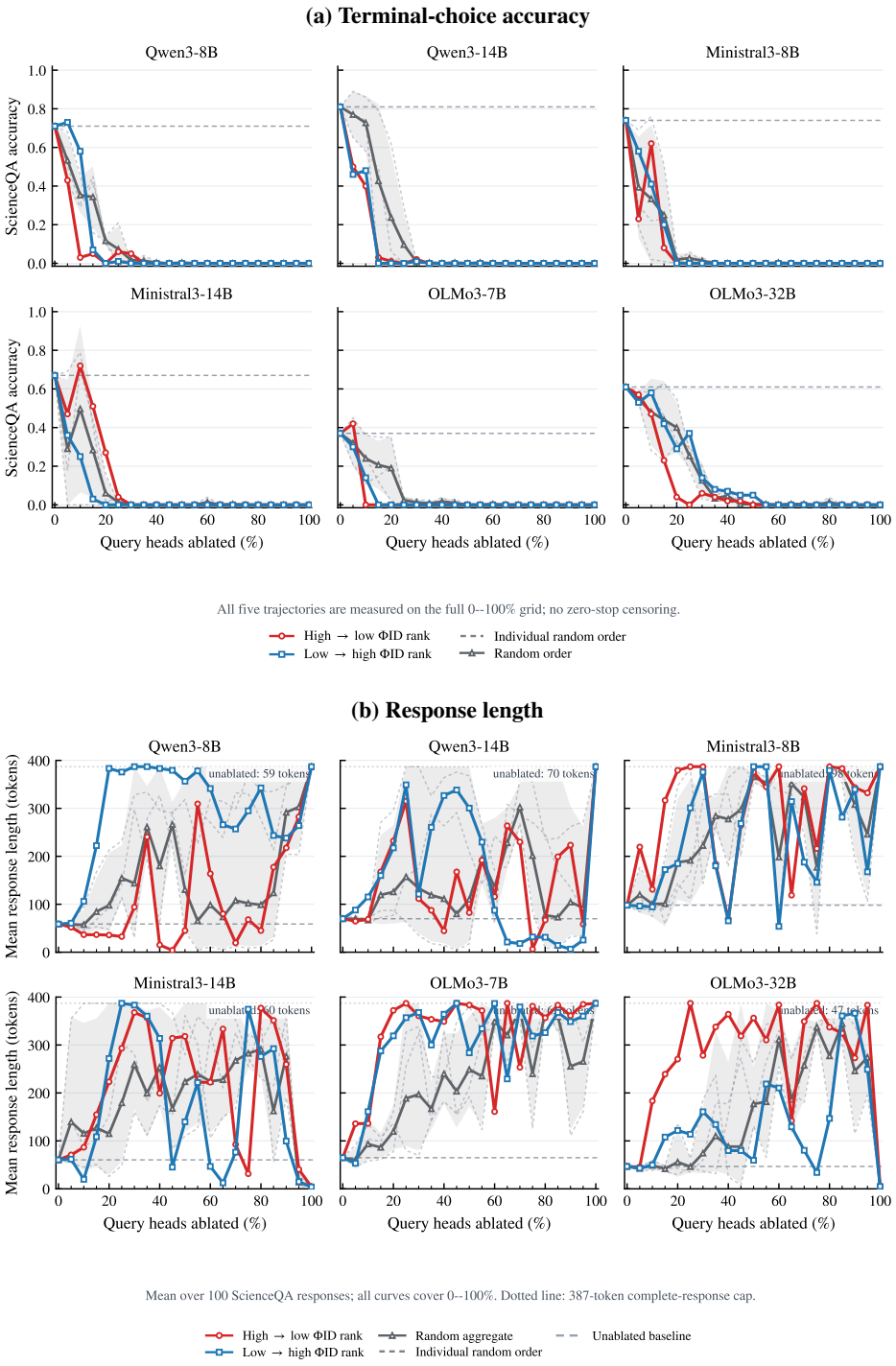


Figure 10: **ScienceQA ablation trajectories.** High-$\mathrm{rank}(S - R)$ ablation typically reduces accuracy earlier and is often accompanied by substantial response-length inflation. Random and low-to-high orders provide the comparison baseline for each model.

E.3 CODE GENERATION

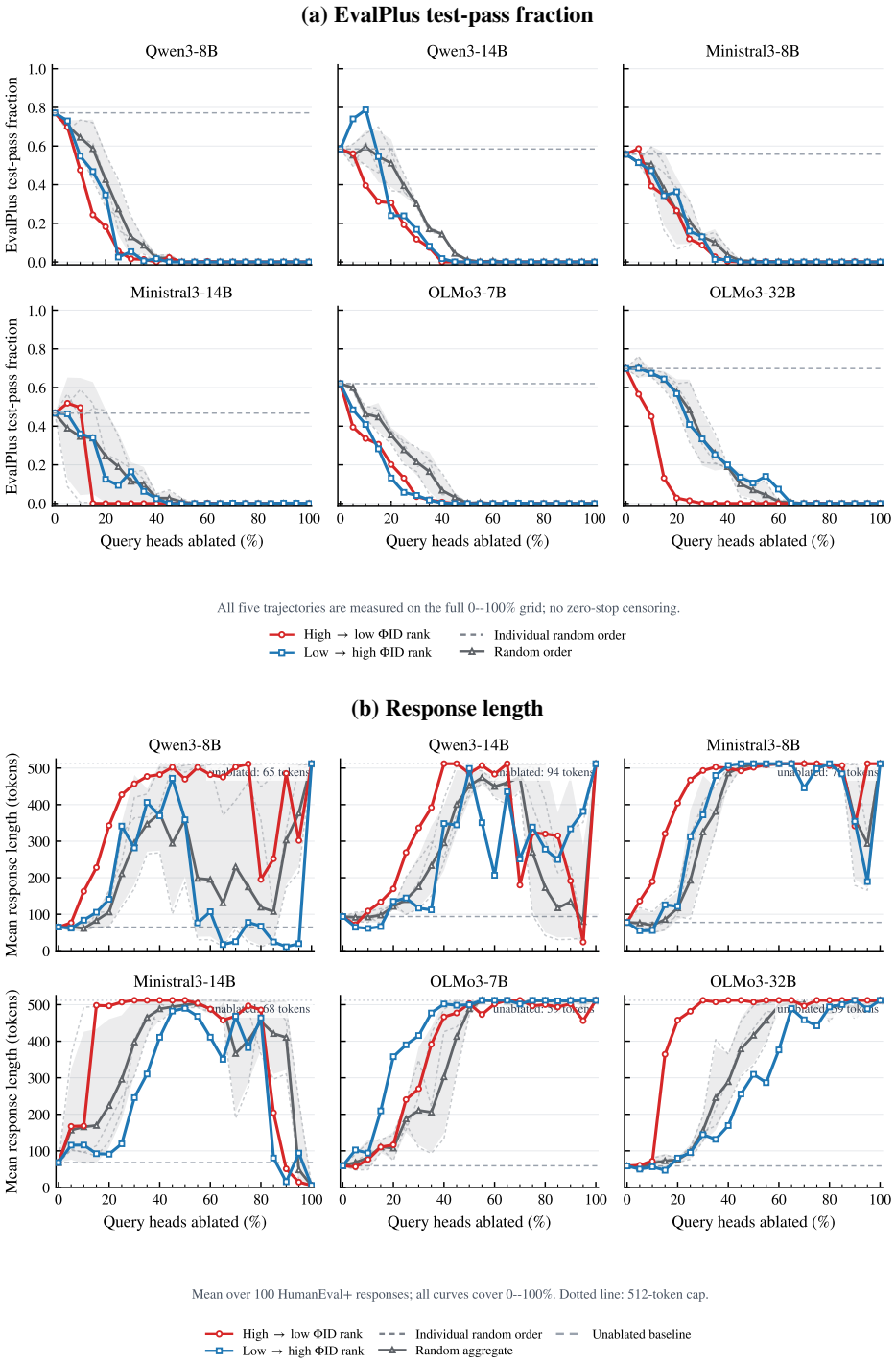


Figure 11: **HumanEval+ ablation trajectories.** Removing high-$\mathrm{rank}(S - R)$ heads first generally reduces the test-pass fraction earlier than random removal. In several models, this degradation is accompanied by much longer generations, showing that score loss and generation dynamics change together.

E.4 LONG-CONTEXT SUMMARIZATION

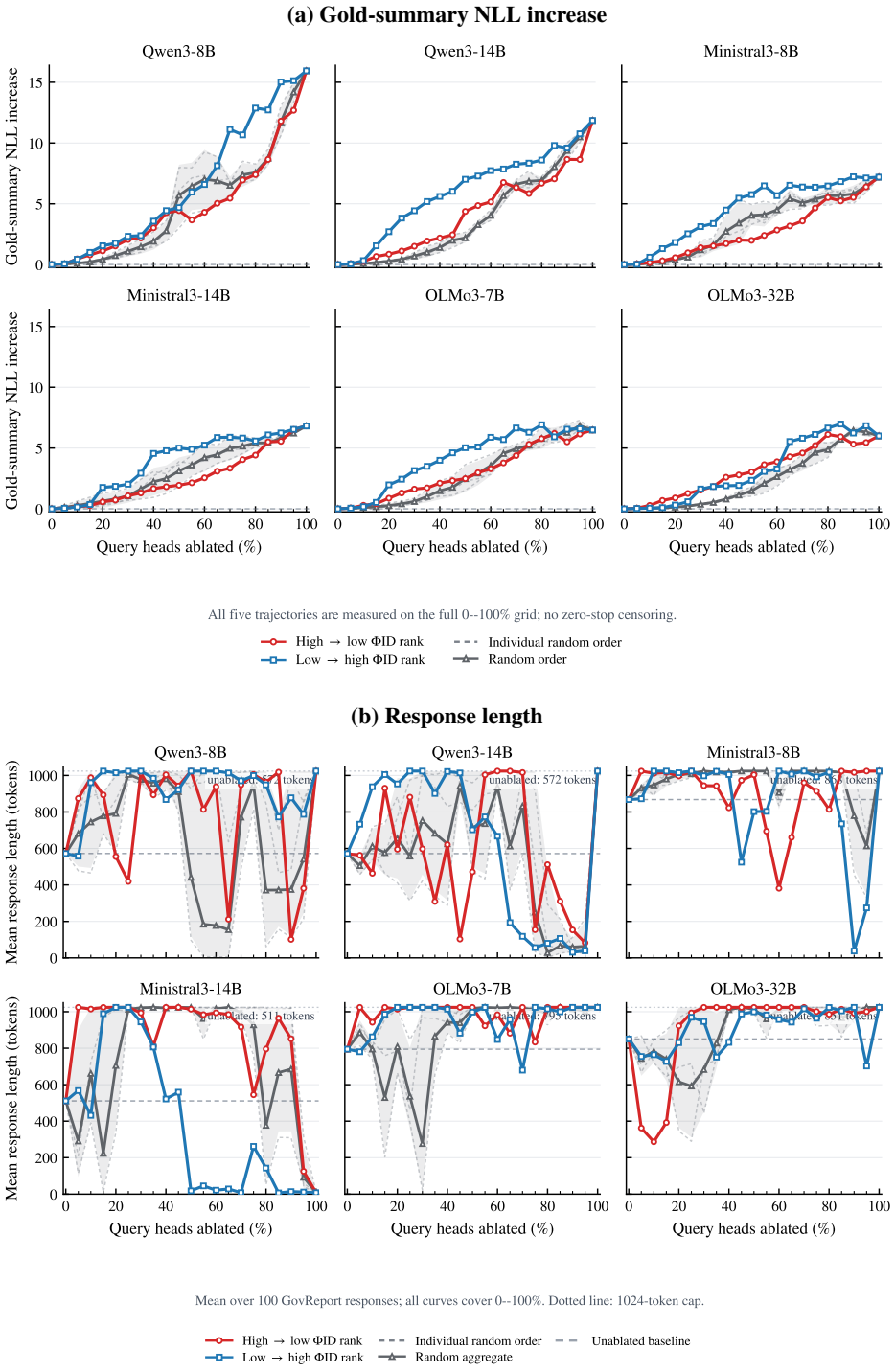


Figure 12: **GovReport ablation trajectories.** Unlike the reasoning tasks, low-to-high $\mathrm{rank}(S - R)$ removal is consistently the most damaging to gold-summary NLL across all six models. At higher ablation levels, the same ordering also tends to shorten the generated summaries.

F CONTROLLED-TASK DETAILS

Matched composition. Numeric prompts apply two or three named non-commuting operations to a one-digit start value. The counterfactual reverses the operation order while preserving the prompt-token multiset, and both target digits are excluded from the worked demonstration. Symbolic prompts contain two disjoint directed chains; the counterfactual rewires the same symbols so that every requested hop has a different target. Each step is queried directly from the original prompt, preventing an earlier answer from being teacher-forced into a later step.

Table 3: Cross-model median formation depths for the final composed relation. Causal onset is based on bidirectional state transfer; explicit onset uses the frozen final norm and output head.

Condition	Causal onset	Explicit onset	Models with explicit onset
Numeric, 2 steps	0.543	0.725	5/6
Numeric, 3 steps	0.500	0.590	6/6
Symbolic, 2 hops	0.470	0.646	6/6
Symbolic, 3 hops	0.470	0.675	6/6

Higher-order intervention cubes. Every frame enumerates all 16 assignments of four independently varied bits and asks for XOR over a designated subset. Complete enumeration makes the Fourier–Walsh coefficients exact for the observed block writes (Walsh, 1923; O'Donnell, 2014; Tsang et al., 2018). Assignment-label permutation retains the set of activation vectors and total variation at each layer while breaking the target mapping, providing a layer-local null for the selected interaction.

G REPRODUCIBILITY STATEMENT

All completed figures derive from immutable manifests recording checkpoint revisions, dataset hashes, specification tokens, seeds, runtime versions, and artifact checksums. Model execution and activation interventions use PyTorch and the Hugging Face Transformers stack; HumanEval+ is pinned to EvalPlus v0.3.1 and test suite v0.1.10 (Liu et al., 2023). Structural plots require all 24 cells to pass finalization; head-ranking and evaluation IDs are disjoint; controlled experiments retain item-level intervention and bootstrap records. Appendix B summarizes protocols, and the anonymous supplement will include plotting code and manifest-linked source tables.