FROM PARAMETERS TO ANSWERS: HOW LLMS RETRIEVE AND USE THEIR INTERNAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) answer questions using pretrained knowledge, but the internal process connecting knowledge retrieval to answer generation remains only partly understood. We find that LLMs use routing information to retrieve relevant pretrained knowledge, then draw on knowledge represented in hidden states to produce an answer. Specifically, parameter routing guides access to pretrained knowledge, while answer-supporting hidden-state knowledge contributes to the answer. Deletion and replacement interventions provide causal evidence for parameter routing and the hidden-state knowledge used to answer. Tracing their effects across layers reveals an overlapping process of knowledge retrieval and use with model- and task-specific timing: parameter routing shapes the knowledge formed in subsequent layers, while hidden-state knowledge supports the answer. In later layers, hidden-state knowledge remains consequential even as parameter-routing deletion causes less answer damage.

1 INTRODUCTION

Large language models (LLMs) (Bai et al., 2023; Brown et al., 2020) acquire extensive knowledge during pretraining and draw on it to answer questions. Yet how they bring the required knowledge into computation and use it across layers remains only partly understood. Prior work has investigated where knowledge is stored in model parameters (Geva et al., 2021; Meng et al., 2022; 2023) and traced information flow during factual retrieval (Geva et al., 2023). Recent studies further examine whether models follow pretrained knowledge or conflicting information in the input prompt (Yu et al., 2023), and how they associate information with the correct entity in context (Feng & Steinhardt, 2024). However, these studies leave unclear what guides knowledge retrieval and how the retrieved knowledge supports the answer across layers. We therefore ask: *how do LLMs progressively retrieve and use their internal knowledge across layers?*

Our first finding is that *LLMs retrieve knowledge through routing and use knowledge represented in hidden states to answer*. **Parameter routing** guides subsequent layers toward relevant pretrained knowledge, while **hidden-state knowledge** supports the answer. Both components are encoded in hidden states. To verify their existence, we delete each component or replace it with the corresponding component from another query, then trace the effects on later hidden states and the answer. Deleting these components makes the model less likely to give the correct answer, while replacing them shifts answer preference toward the other query's answer.

Our second finding is that *parameter routing and hidden-state knowledge vary in their influence on later computation and answers across layers*. Early layers build representations of the query. As computation proceeds, parameter routing guides access to relevant pretrained knowledge, which develops into answer-supporting knowledge in hidden states. These processes overlap, with timing that varies across models and tasks. In later layers, parameter-routing deletion causes less answer damage, while deleting hidden-state knowledge continues to damage the answer. We trace this progression by deleting or replacing both components at different layers and measuring their effects on later hidden states and answers.

In summary, our contributions are twofold. (1) Using deletion and replacement interventions, we identify parameter routing and hidden-state knowledge as distinct functional components. (2) We reveal how their contributions to knowledge retrieval and use change across layers.

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.

2 Related Work

Parametric knowledge representation. Prior work shows that pretrained language models can answer factual queries without fine-tuning and identifies neurons associated with particular facts (Petroni et al., 2019; Dai et al., 2022). Researchers also characterize feed-forward layers as associative memories and develop methods that locate and edit specific knowledge associations without retraining the entire model (Geva et al., 2021; Meng et al., 2022; 2023). Localization and effective editing need not select the same layers (Hase et al., 2023). Our question instead concerns the changing roles of request information and answer-supporting content during the unedited model's forward computation.

Internal computation of knowledge retrieval. Another line of work examines how language models retrieve and use pretrained knowledge during inference. These studies characterize how model components transform query information into knowledge-based predictions and decompose retrieval into functional elements such as an input, an operation, and a returned value (Geva et al., 2023; Hernandez et al., 2024; Wang & Xu, 2025). Related analyses identify heads that promote pretrained versus in-context answers and inspect what information hidden states make readable (Yu et al., 2023; Ghandeharioun et al., 2024). Building on this functional view, we compare the layer-wise consequences of separately changing a fitted request component and fitted answer-supporting content, measuring both final answers and later content. This comparison asks when the remaining computation depends on each measured component, rather than treating readable information as evidence of its causal use.

Causal analysis of internal knowledge use. Mechanistic studies intervene on hidden states to identify internal information that influences knowledge-based predictions and trace its propagation through the model (Wang et al., 2024; Hochman et al., 2026). Intervention results can depend on the chosen control and score, and a successful subspace intervention need not recover the model's natural computation (Zhang & Nanda, 2024; Makelov et al., 2024). An effect on the answer therefore does not by itself identify the role that the affected information plays: the state may carry answer-supporting content or guide subsequent layers toward that content. Our work distinguishes these roles and traces how their causal contributions change across layers.

3 Retrieval of Knowledge from Model Parameters

In this section, we investigate how LLMs access knowledge stored in their parameters. We identify *parameter routing* in hidden states that guides access to relevant pretrained knowledge, examine how its influence changes across layers, and check whether the edits could instead alter factual information already present at the edited layer.

3.1 Evidence for Parameter Routing

We test whether parameter routing exists in hidden states and guides access to the knowledge needed to answer. If a hidden-state component performs this function, deleting it should weaken the model's preference for the correct answer. Replacing it with the corresponding component from another question should shift later hidden-state knowledge and answer preference toward that question's answer. We test these predictions through deletion and replacement interventions.

We use countries from Africa, Asia, and Europe, always pairing countries from different continents. For each pair, one question wording constructs a candidate direction, and two other wordings evaluate it (Appendix B). For example, for the held-out Angola–Japan pair, the construction wording asks "Which continent is Angola located in?" and the same question about Japan. We input these two questions separately into Qwen-2.5-3B-Instruct (Qwen Team, 2024). At layer ℓ, we subtract the second question's final input-token hidden state from the first's and denote the difference by Δ^ℓ. It captures how the recorded state changes when the question requires knowledge about a different country. As the paired questions have different continent answers, Δ^ℓ may also contain differences in hidden-state knowledge about those continents. Directly intervening on Δ^ℓ could therefore change the represented knowledge rather than how the model accesses knowledge. To control for this possibility, we estimate a two-dimensional subspace associated with differences among three

continents using a separate set of questions. At each layer, we average their hidden states by the correct continent, subtract the overall mean, and apply singular value decomposition to the resulting mean differences (Shlens, 2014; Halko et al., 2011). Let $\mathbf{P}^\ell$ denote the orthogonal projection onto this subspace. Then $\Delta^\ell_{\text{cont}} = \mathbf{P}^\ell \Delta^\ell$ is the part of the paired difference in the fitted two-dimensional space.

We remove $\Delta^\ell_{\text{cont}}$ and normalize the remainder:

$$\mathbf{r}^\ell = \texttt{Normalize}\big(\Delta^\ell - \Delta^\ell_{\text{cont}}\big)\,. \tag{1}$$

Here $\texttt{Normalize}(\cdot)$ is a normalization function (Trefethen & Bau, 1997). $\mathbf{r}^\ell$ represents the normalized paired difference after removing its fitted continent-knowledge component. Each country pair has its own $\mathbf{r}^\ell$, which we keep fixed when testing the other two wordings. We use it to identify the candidate parameter-routing component to delete or replace.

For deletion, we edit each evaluation question in turn, using the other question in its pair as the counterpart.

Proposition 1. *For paired evaluation questions q and $\bar{q}$ at layer ℓ, let $\mathbf{h}^\ell_q$ and $\mathbf{h}^\ell_{\bar{q}}$ be their final input-token hidden states. Define*

$$\alpha^\ell_q = (\mathbf{r}^\ell)^\top \left(\mathbf{h}^\ell_q - \tfrac{1}{2}(\mathbf{h}^\ell_q + \mathbf{h}^\ell_{\bar{q}})\right). \tag{2}$$

This scalar is the signed amount of $\mathbf{r}^\ell$ in $\mathbf{h}^\ell_q$ relative to the midpoint of the paired states. Deletion removes that amount:

$$\mathbf{h}^\ell_{q,\text{del}} = \mathbf{h}^\ell_q - \alpha^\ell_q \mathbf{r}^\ell. \tag{3}$$

Then $\mathbf{h}^\ell_{q,\text{del}}$ is the hidden state with the identified parameter-routing component removed, while all orthogonal components remain unchanged.

For replacement, we substitute the corresponding component from $\bar{q}$ into the hidden state for q, as formalized below.

Proposition 2. *For paired evaluation questions q and $\bar{q}$ at layer ℓ, define*

$$\mathbf{h}^\ell_{q,\text{rep}} = \mathbf{h}^\ell_q - 2\alpha^\ell_q \mathbf{r}^\ell. \tag{4}$$

Then $\mathbf{h}^\ell_{q,\text{rep}}$ is the hidden state obtained by replacing the parameter-routing component of $\mathbf{h}^\ell_q$ with that of $\mathbf{h}^\ell_{\bar{q}}$, while leaving all orthogonal components unchanged.

Appendix B.5 proves Propositions 1 and 2. We pass each edited hidden state through the remaining layers and evaluate both interventions using the same protocol.

For Table 1, we intervene at layer 32 and measure the later state at layer 36; three screening pairs select these layers, and 24 held-out pairs provide the results. The projected mean state for each continent at layer 36 serves as that continent's reference. The fitted-state scores $s_{q,1}$ and $s_{\bar{q},1}$ are the negative squared distances from the projected later state to the references for the correct continents of q and $\bar{q}$. The answer scores $s_{q,2}$ and $s_{\bar{q},2}$ are the mean token log probabilities of their corresponding continent answers. All four scores come from the same run for q; $\bar{q}$ supplies the competing continent. A larger value indicates stronger support for that score's continent. Because activation-intervention conclusions can depend on the corruption and score used (Zhang & Nanda, 2024), each targeted edit is compared with eight Gaussian edits matched to its vector length; deletion and replacement use separate random controls. Appendix B.6 gives the detailed calculations. For the layer-wise analysis, we use the two differences $s_1 = s_{q,1} - s_{\bar{q},1}$ and $s_2 = s_{q,2} - s_{\bar{q},2}$.

Findings. Table 1 reports all four scores under deletion, replacement, and their separately matched random interventions. Deletion reduces the answer-preference margin s_2 beyond matched random edits, but its later fitted-state effect is inconclusive because its interval spans zero (Appendix B.7). Replacement decreases both scores for q while increasing both scores for $\bar{q}$ relative to its matched random intervention. Its effects on the later fitted state and answer both exceed the matched controls, showing a shift toward the knowledge requested by $\bar{q}$. Together, replacement's effect on later measured knowledge and the answer, plus deletion's answer effect, support a routing contribution of the fitted coordinate. We required both interventions to exceed random controls on both outcomes; this joint criterion fails because deletion's later-knowledge interval includes zero.

Table 1: Qwen country–continent scores at $32 \rightarrow 36$. State scores are negative squared distances to continent references, shown in thousands; answer scores are mean token log probabilities. Both answers are scored from q's run.

Condition	State score		Answer score	
	$s_{q,1}$	$s_{\bar{q},1}$	$s_{q,2}$	$s_{\bar{q},2}$
No intervention	−2.08	−31.39	−0.004	−13.696
Random for deletion	−2.30	−30.47	−0.005	−13.325
Deletion	−2.31	−29.08	−0.008	−10.765
Random for replacement	−2.87	−28.16	−0.024	−12.368
Replacement	−4.42	−23.17	−0.643	−5.646

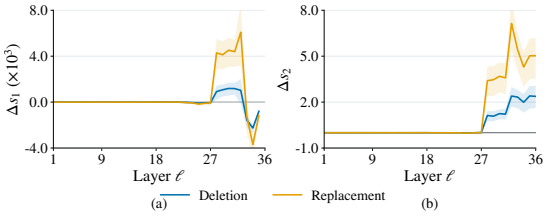


Figure 1: Qwen effects beyond matched random edits: (a) Effects on later hidden-state knowledge, measured by Δs_1 at layer 36 after edits at layers 1–35; (b) Effects on answer preference, measured by Δs_2 after edits at layers 1–36. Shading gives pointwise 95% country-pair bootstrap intervals.

3.2 Layer-wise Development of Parameter Routing

Experimental setup. We next examine how the influence of parameter routing changes across layers. We apply deletion and replacement at different layers and measure their effects on later hidden-state knowledge and answer preference.

For each outcome s_1 or s_2, Δs_1 or Δs_2 is its mean decrease from the no-intervention condition under the targeted edit, minus the largest corresponding mean decrease among eight matched random edits. Table 1 instead shows their average raw scores. We measure s_1 at layer 36 throughout. Appendix B.6 provides the detailed calculation and aggregation procedures.

Findings. We report Δs_1 and Δs_2 in Fig. 1(a) and (b), respectively. In early layers, neither deletion nor replacement has a clear excess effect over its matched random edits. Around layers 28–32, replacement increasingly decreases both the later fitted-state margin and the answer margin; deletion's answer effect also rises. At the fixed layer-32 comparison, however, deletion's later fitted-state effect remains inconclusive because its interval spans zero (Appendix B.7). At edit layers 34–35, both interventions decrease the later fitted-state margin less than their strongest matched random edits, so their Δs_1 values turn negative; their answer effects remain positive. These results identify a late range where replacement changes both measurements, but do not establish why the two effects diverge near the output.

3.3 Additional Analysis

(1) Measured content at the edited layer. To check whether deleting or replacing the component along $\mathbf{r}^\ell$ directly changes the continent information measured at that layer, we project the changes actually written after floating-point rounding onto the separately fitted two-dimensional continent-content space. None of the recorded deletion or replacement writes fails the registered residual check, and the largest projected fraction is below one millionth of its write length; this preserves the measured linear component, not necessarily fact-related information outside that space (Appendix B.8).

(2) Wording and edit length. To test whether the effects are specific to one wording or edit length, we construct a separate direction for each of 24 country pairs excluded from content fitting and layer screening using one wording, then test that pair under two other wordings against equal-length random edits. Deletion lowers the requested answer's score advantage, and replacement shifts both that advantage and later fitted knowledge toward the paired answer beyond these controls; no direction is shared across pairs (Appendices B and B.7).

(3) Random-control sensitivity. To check whether the uncertain deletion effect on later fitted knowledge changes with the random controls, we leave the unmodified and targeted runs fixed and recompute the comparison after pooling eight additional controls with the original eight. Its interval still spans zero, while the replacement effect remains positive; the joint criterion requiring both edits to affect both outcomes still fails (Appendix H.3).

(4) Independent model reload. To check computational reproducibility, we reload the model and repeat the frozen comparisons, with one repeated shard running on different hardware. The selected layers, pass/fail decisions, and rounded estimates agree although individual floating-point records are not bitwise identical; the repeat is not an additional statistical sample (Appendix B.8).

(5) Request before the country. To test whether a request can affect factual answering before the current country is seen, a separate Qwen task places the relation first, as in *Relation to answer: continent. Country entity: Kenya. Provide only the corresponding value.*, and deletes its natural continent-versus-capital coordinate fitted from other wordings at that earlier position. The relation-word state is identical across country suffixes; at layer 16, targeted deletion lowers the answer advantage for all 16 questions and by 10.51 on average, compared with 0.92 for the country-position edit and 0.88 for the strongest equal-length random edit. This supports an early request contribution, not identity with the pair-specific $\mathbf{r}^\ell$ or a measured later-knowledge path (Appendix B.9).

4 Formation and Use of Hidden-State Knowledge

Having found that replacing the fitted parameter-routing component shifts later measured knowledge and the answer, we next ask whether factual content measured in hidden states itself supports the answer. We first test deletion and paired-reference replacement at fixed layers, then trace their effects across layers and answer formats, and finally examine measurement alignment and transfer across natural and arbitrary answer forms.

4.1 Evidence for Answer-Supporting Knowledge

If factual content represented in a hidden state supports the answer, we should observe two patterns. First, removing the fitted content should make the correct answer less likely. Second, paired-reference replacement should make the paired answer more likely. We test these two hypotheses separately before examining their layerwise development.

4.1.1 Deletion Intervention

Experimental setup. We use paired-country questions that list two countries but request the continent of only one. For example, q asks for Kenya's continent while $\bar{q}$ changes only the requested position and asks for China's; neither input supplies a continent. We evaluate Qwen, Llama, and Gemma (Qwen Team, 2024; Dubey et al., 2024; Gemma Team, 2025) on country-disjoint fitting, screening, and validation groups. Appendix C gives the complete prompts, model versions, answer formats, and partitions.

Let $\mathbf{h}_q^\ell$ be the final input-token state for question q at layer ℓ. For each model, layer, and answer format, we average the fitting states by the requested continent. After subtracting the overall fitting mean, the three continent means span a two-dimensional fitted space. Let $\mathbf{P}^\ell$ project onto this space, $\boldsymbol{\mu}^\ell$ be the overall fitting mean, and $\boldsymbol{\mu}_k^\ell$ the mean for continent k. We define

$$\mathbf{z}_q^\ell = \mathbf{P}^\ell(\mathbf{h}_q^\ell - \boldsymbol{\mu}^\ell), \qquad \boldsymbol{\kappa}_k^\ell = \mathbf{P}^\ell(\boldsymbol{\mu}_k^\ell - \boldsymbol{\mu}^\ell). \tag{5}$$

Here $\mathbf{z}_q^\ell$ is the centered state projected into the fitted space, and $\boldsymbol{\kappa}_k^\ell$ is continent k's reference in that space. These are analysis interfaces, not assumed complete native representations. Projection-based erasure has been used to remove measured properties from representations (Ravfogel et al., 2020); testing whether such information affects behavior requires an intervention (Elazar et al., 2021). The requested-minus-paired answer score s_2 compares the probability assigned to the requested continent with the probability assigned to the other listed country's continent. We compare how much each targeted edit and its equal-length random controls decrease s_2; Appendix C gives the exact calculation.

Lemma 1. *For question q at layer ℓ, the edit*

$$\mathbf{h}_{q,\mathrm{del}}^\ell = \mathbf{h}_q^\ell - \mathbf{z}_q^\ell \tag{6}$$

deletes the fitted continent-related component from the state for q while leaving every component orthogonal to the fitted space unchanged.

Appendix C.6 proves Lemma 1. We apply the deletion at one decoder-block output, run the remaining layers, and test whether the requested-minus-paired answer score decreases.

Separate screening questions fix one comparison layer per model before evaluation on two held-out groups. Appendices C and D give the fitting, layer-selection, control, and aggregation rules.

Findings. At the fixed comparisons, deletion decreases the requested-minus-paired answer score more than the matched random controls in both held-out groups. Because the component is fitted across distinct countries that share an answer, this result links the cross-country continent-related component to the answer rather than to one fitting country's identity.

4.1.2 Paired-Reference Replacement Intervention

Experimental setup. Deletion does not test whether the fitted space points to a specific alternative fact. To test the second hypothesis, let k_q^+ and k_q^- be the requested and paired continents. We use the same comparison layers fixed by the deletion screening before any paired-reference replacement result is examined (Appendix D).

Lemma 2. *For question q at layer ℓ, the edit*

$$\mathbf{h}^{\ell}_{q,\mathrm{rep}} = \mathbf{h}^{\ell}_{q} + \boldsymbol{\kappa}^{\ell}_{k_q^-} - \boldsymbol{\kappa}^{\ell}_{k_q^+} \tag{7}$$

replaces the requested continent reference with the paired continent reference in the fitted-space component while leaving every component orthogonal to the fitted space unchanged.

Appendix C.6 proves Lemma 2. We use the edited state to test whether the requested-minus-paired answer score decreases more than under matched random edits. Appendices C, D, and F give the complete controls and aggregation rules.

Findings. Paired-reference replacement supplies the complementary directional result: in all three models, it decreases the requested-minus-paired answer score more than its matched random controls in both validation groups (Table 2).

Table 2: Fixed-layer decreases in the requested-minus-paired answer score. Registered layers are Qwen 34, Llama 28, and Gemma 33.

Model	Intervention	Condition	Validation A	Validation B
Qwen	Deletion	Random	0.96	0.79
		Targeted	8.83	8.19
	Replacement	Random	3.92	4.13
		Targeted	25.88	25.02
Llama	Deletion	Random	1.76	1.98
		Targeted	18.10	17.96
	Replacement	Random	3.70	4.22
		Targeted	32.27	32.14
Gemma	Deletion	Random	3.92	3.75
		Targeted	24.17	23.42
	Replacement	Random	13.15	11.50
		Targeted	62.23	60.79

Together, deletion shows that the fitted content makes a causal contribution to the requested answer, while replacement shows that moving the fitted reference toward the paired continent redirects the answer. The claims concern this fitted two-dimensional space; they do not imply that it contains all answer-relevant information.

4.2 Layer-wise Development of Answer-Supporting Knowledge

Experimental setup. We next trace when deleting the fitted component begins to affect the answer across layers and answer formats. We analyze distinguishability separately because a projected

state can be closer to the correct reference before deleting it causes answer damage beyond random controls. For question q at layer ℓ, let k_q^+ and k_q^- be the requested and paired continents. We define

$$G_q^\ell = \|\mathbf{z}_q^\ell - \boldsymbol{\kappa}_{k_q^-}^\ell\|_2^2 - \|\mathbf{z}_q^\ell - \boldsymbol{\kappa}_{k_q^+}^\ell\|_2^2.$$

A positive G_q^ℓ means that the projected state is closer to the requested reference. The fraction of questions with positive G_q^ℓ measures distinguishability, not causal use. The causal curves report $E_{\text{know}}^{\ell,\text{ans}}$, the decrease in s_2 after fitted-knowledge deletion minus the largest decrease under matched random edits. Appendix D gives the complete profiles and aggregation rules.

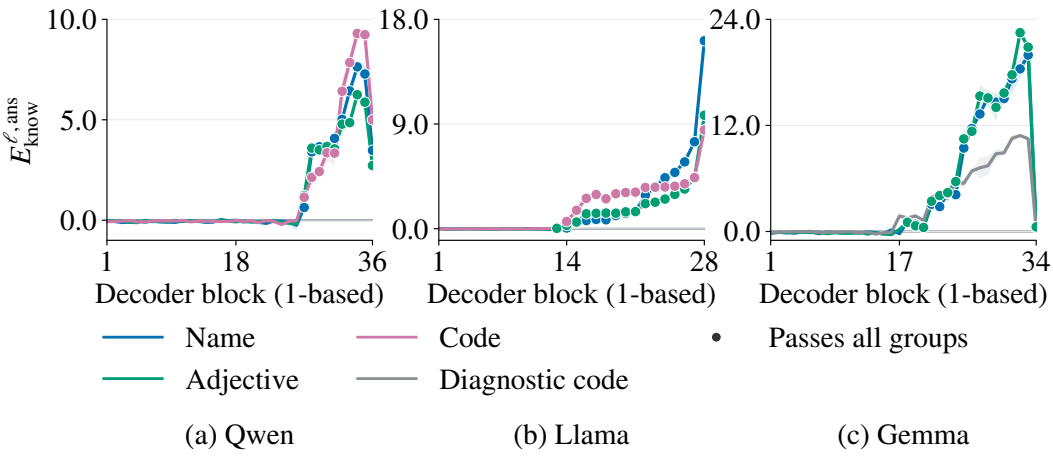


Figure 2: **Layerwise answer effects of fitted-knowledge deletion across answer formats.** $E_{\text{know}}^{\ell,\text{ans}}$ is the random-corrected decrease in the requested-minus-paired answer score at layer ℓ; filled dots use each curve's color and mark layers passing screening and both validation groups. Gray denotes Gemma's diagnostic code condition.

Findings. Figure 2 reports the causal answer effect after deleting the fitted component at each layer. Lines average the two validation groups, bands span their values, and filled dots additionally require positive screening and both validation effects. For continent names and adjectives, passing layers occur predominantly late in all three models. Code answers show the same late pattern in Qwen and Llama; Gemma's code result remains diagnostic because its unmodified answers are not all correct in one validation group. The fraction of questions with positive G_q^ℓ rises before deletion passes all groups (Appendix D, Figs. 4, 6, and 7), showing that readable content can precede detectable causal use. Appendix F reports the complementary replacement scans.

4.3 Additional Analysis

(1) Answer forms. To test whether the fitted content applies beyond continent-name answers, we use the continent-name-fitted reference change on adjective and arbitrary-code questions without refitting it; the code prompt supplies the continent-to-code mapping but not either country's continent. The continent-name-fitted change shifts adjective answers in all three models, whereas among baseline-qualified code mappings it passes for only one Qwen mapping and none in Llama or Gemma, so its measured effect covers another natural expression rather than every output mapping (Appendix F).

(2) Layerwise checks. To examine when the fitted content becomes readable and affects answers, we compare the unedited state's correct-reference proximity with the answer change after deletion at each layer and scan paired-reference replacement across the same answer formats. The unedited projected state can be closer to the requested continent reference before deletion changes the answer beyond matched random edits; replacement effects are predominantly late, but that scan reuses the original validation questions and is descriptive (Appendices D and F).

(3) Fitting and random controls. To check whether the deletion result relies on particular fitting questions or random edits, we compare the full fit with three fits that each omit one country pair,

under three random-control seeds while holding both validation groups fixed. At every listed comparison layer, all four fits retain a positive random-corrected content-deletion effect across these seeds and groups, so the measured deletion effect is not tied to one fitting pair or random draw (Appendix F).

(4) New countries and state coordinates. The original profiles record the returned final-layer state after normalization but edit the decoder-block output before it; we therefore keep the original fitting questions, exclude 24 new country pairs from fitting, and test eight variants per pair that change country order, wording, and requested country while fitting, measuring, and editing at decoder-block outputs. All unmodified questions prefer the correct answer, and later-layer deletion lowers its score advantage beyond matched random edits with country-pair bootstrap intervals above zero in all three models; because both the country set and coordinate convention change, this does not isolate the coordinate difference (Appendices D and E).

Sec. 5 uses these new pairs for its fixed earlier-to-later comparison.

5 How Internal Knowledge Is Progressively Retrieved and Used

Having established parameter routing and answer-supporting hidden-state knowledge separately, we now compare how their causal effects change from early to late layers. Prior work finds factual retrieval to involve distributed, non-contiguous components (Hochman et al., 2026); our comparison measures the two fitted components on the same questions.

5.1 Layer-wise Development of Retrieval and Use

Experimental setup. We use the paired-country continent-name questions evaluated with Qwen, Llama, and Gemma. At each decoder block, we apply the paired-task parameter-routing deletion corresponding to Sec. 3 and the fitted-knowledge deletion from Sec. 4. Both operations modify the final input-token state, use the same requested-minus-paired answer score, and are compared with equal-length random changes. We write $E^{\ell,\mathrm{ans}}_{\mathrm{par}}$ and $E^{\ell,\mathrm{ans}}_{\mathrm{know}}$ for their random-corrected answer effects. Appendix D.2 gives the paired-task routing construction, and Appendix E gives the fixed groups and aggregation rules.

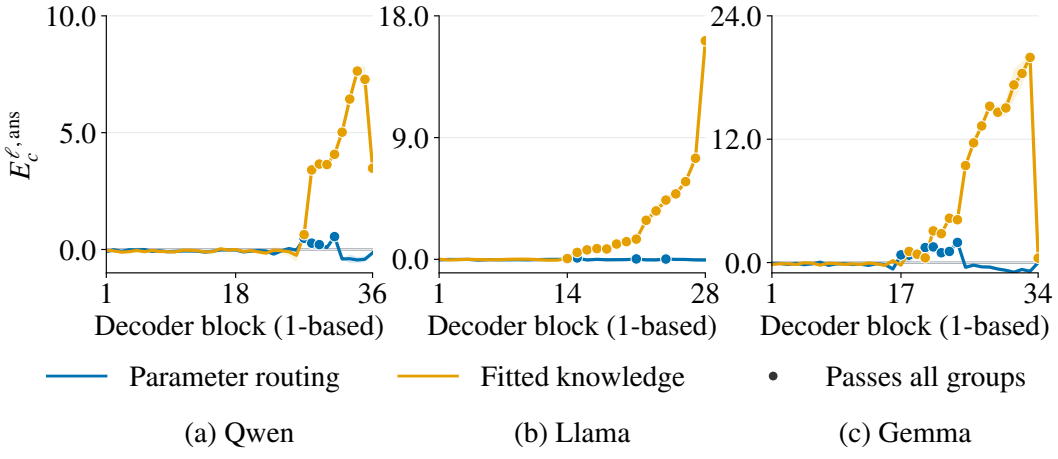


Figure 3: **Layerwise parameter-routing and fitted-knowledge deletion effects for continent-name questions.** $E^{\ell,\mathrm{ans}}_c$ is the random-corrected decrease in the requested-minus-paired answer score after deleting component c at layer ℓ; lines average the two validation groups, bands span their values, and filled dots use each curve's color and require positive screening and both validation effects.

Findings. Figure 3 shows three consistent parts of the progression. In early layers, neither deletion has a stable answer effect. In transition layers, parameter-routing and fitted-knowledge deletion begin to exceed random controls at overlapping but model-specific depths: they begin together in

Qwen, parameter routing begins slightly earlier in Gemma, and fitted knowledge is more continuous in Llama. In late layers, fitted-knowledge deletion remains consequential after parameter-routing deletion stops passing the registered criterion. The results therefore support overlapping retrieval and use rather than one universal boundary shared by all models.

Table 3: Model-specific layer ranges passing screening and both validation groups.

Model	Parameter routing	Fitted knowledge
Qwen	27–29, 31	27–36
Llama	15, 21, 24	14–28
Gemma	17, 19–24	18–34

Table 3 makes the model differences explicit while preserving the common late pattern. On 24 additional country pairs, the fixed earlier-to-later comparison gives the same directional result: parameter-routing deletion causes less answer damage in the later sets, whereas late fitted-knowledge deletion causes more damage than late parameter-routing deletion. Each pair contributes several wordings and request variants, but uncertainty resamples country pairs, not the correlated variants; the intervals for both contrasts exclude zero in all three models. Opposite requests remain separable along the late fitted request direction, so reduced deletion damage does not mean that request information has disappeared.

A single late deletion might appear weak because later blocks rebuild the measured routing component. In Qwen, we therefore repeat deletion at every late block and match random changes to the actual written lengths. Relative to deletion at the first late block alone, the added answer loss is only 0.21% of the unmodified answer margin, with a one-sided 95% upper bound of 0.45% below the prespecified 10% bound; repeated fitted-content deletion remains consequential. This check concerns the measured global routing direction, not all possible request information (Appendix E).

The paired-task direction here is fitted across questions, whereas Sec. 3 constructs a separate direction for each country pair. Holding Qwen and the paired task fixed, a common-protocol check finds positive late answer damage from the pair-conditioned deletion under either centering choice, but not from deletion of the fitted global direction. At the same source and receiving layers with the same fitted-content measurement and fit center, the pair-conditioned edit also causes a positive later-knowledge loss, whereas the global edit's interval spans zero. The actual write lengths differ, so these comparisons do not isolate direction alone; the late-layer conclusion applies to the measured global edit (Appendix H).

Together, the figure and table show overlapping layerwise effects: deleting the measured paired-task request component causes more answer-score loss than matched random edits at selected middle layers, while deleting fitted factual content continues to cause extra loss later. The exact onset and persistence vary by model and measurement interface, so the evidence does not establish universal processing stages or aligned layer numbers.

6 Conclusion

Our experiments distinguish request information that guides access to parametric knowledge from factual content in hidden states that supports the answer. Deletion and replacement interventions show that both measured components make causal, model- and task-dependent contributions. Their layerwise effects overlap rather than forming universal stages: parameter-routing deletion becomes less consequential in later layers, while deleting fitted factual content continues to damage the answer. The fitted content transfers between natural noun and adjective expressions but generally not to arbitrary codes, further separating reusable factual content from one output-independent representation. These results characterize changing dependence on specified hidden-state components, not a claim that the fitted measurements exhaust the model's retrieval mechanism.

AI USE STATEMENT

We used generative AI tools to assist with literature organization, experimental and figure-generation code, data analysis, and manuscript drafting and revision. The authors reviewed all AI-assisted text and code, verified reported measurements against saved result files, and compiled and visually inspected the final paper. We take responsibility for the complete content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Each result section states the question, fitted measurement object, intervention, control, score, and decision rule. Appendix C defines the shared paired-country protocol. Appendices D–H provide the remaining fitting, centering, calibration, repeated-intervention, direct-selection, and robustness details. Figure directories contain plotted data, file hashes, and plotting programs. An anonymized code-only archive for these experiments is available at `https://anonymous.4open.science/r/rrp-pretrain-4426/README.md`.

Frozen records retain model and tokenizer identities, rendered prompts, intervention positions, per-question outputs, and fresh-load comparisons. Compact evidence for the base-model and instruction-model capital-selection experiments is included under `evidence/object_selection_capital/`; the natural-route timing and complementary three-model reanalyses are under `evidence/route_strength_timing/`. Both directories include source hashes. Fresh loads check computational reproducibility and are never counted as additional statistical samples.

The original paired-task bands and the complementary three-model trajectory figure show the range of two validation splits. Country-pair bootstrap intervals are used in the natural single-country, expanded-country, and common-protocol Qwen analyses; the direct capital-selection experiment reports fixed split-level question fractions. The natural-question, common-protocol Qwen, and direct capital-selection experiments use float32 and verify the vector actually written. Whenever a screening gate fails, the paper states that held-out intervention or two-candidate restoration was not run rather than treating the absent records as missing data.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, 2020.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581. URL `https://aclanthology.org/2022.acl-long.581/`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986. doi: 10.1214/ss/1177013815.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL `https://aclanthology.org/2021.tacl-1.10/`.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *ICLR*, 2024.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *EMNLP*, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, pp. 12216–12235, 2023.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490, 2024.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *NeurIPS*, volume 36, 2023.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *ICLR*, 2024.

Hail Hochman, Natalie Shapira, and Yoav Goldberg. Factual retrieval in LLMs is a redundant, distributed and non-contiguous process. In *ACL*, pp. 46747–46768, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *ICLR*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, volume 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *ICLR*, 2023.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pp. 2463–2473, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014. URL https://arxiv.org/abs/1404.1100.

Lloyd N. Trefethen and III Bau, David. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

Zijian Wang and Chang Xu. Functional abstraction of knowledge recall in large language models. *arXiv preprint arXiv:2504.14496*, 2025.

Zijian Wang, Britney Whyte, and Chang Xu. Locating and extracting relational concepts in large language models. In *Findings of ACL*, pp. 4818–4832, 2024.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In *EMNLP*, pp. 9924–9959, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

A LIMITATIONS

The results establish functional distinctions under specific tasks and fitted measurements. Their scope is limited by the knowledge relations and models tested, the information captured by the linear measurements, and the interventions used to assess dependence. These limitations qualify how far the findings generalize without changing the within-task results.

The main evidence concerns country–continent associations. Paired-country experiments cover three instruction-tuned models, whereas the natural single-country intervention is complete only for Qwen. Nouns, adjectives, and arbitrary codes change how one association is expressed; they are not three knowledge domains. The expanded evaluation uses 24 held-out country pairs, but all fitted directions and content spaces are estimated from small country sets. These choices limit population, relation, and fitting-sample generality.

The natural-route timing analysis reuses frozen Qwen validation states after the original experiment and therefore supplies a diagnostic onset rather than a prospectively registered change point. Only Qwen has the full natural single-country states needed to divide the pair-conditioned coefficient by the whole-state scale. The three-model paired-task analysis uses a different global direction and normalizes by fitted request separation; it can reveal complementary model-specific trajectories but is not a direct Llama/Gemma replication of the Qwen timing result.

The country–continent prompts do not state the target association. Supplied markers appear only in the auxiliary selection task, and code prompts provide only a continent-to-code output map. Nevertheless, the evaluated instruction checkpoints combine pretraining with later training. The experiments study how the released weights use these associations; they do not assign each fact exclusively to pretraining rather than instruction tuning or another training stage.

Every named route and content component is a fitted linear measurement at the final input position. The two-dimensional continent space is not an exhaustive account of how a model represents the fact, and projecting it out does not remove every possible encoding. Likewise, the pair-conditioned natural-question direction is built separately for each country pair. Its pointing score tests transfer from one wording to two others within that pair; it is not a classifier trained on old countries and applied independently to unseen identities.

The common-protocol Qwen comparison varies request-direction construction under matched model and task conditions, but each deletion removes its own natural projection. Direction and intervention length therefore change together, as quantified in Sec. H.1; we did not compare the two orientations at one common targeted length. The observed earlier-to-later decline in answer damage concerns deletion of the fitted paired-task global request direction, not all request representations, and fitted content is not shown to be its unique mediator. In the natural single-country experiment, deletion-based loss of later knowledge also remains inconclusive at the fixed layer pair after pooling sixteen random controls; only replacement gives the confirmed effect there. Neither comparison establishes a universal transfer of causal control from all request information to content.

Noun-to-adjective content transfer does not establish an expression-independent representation: most baseline-qualified arbitrary code maps fail to transfer. The supported claim is reuse across the tested natural expressions of a continent, not one output-independent content component shared by every answer mapping.

B NATURAL-QUESTION PARAMETER-ROUTING PROTOCOL AND RESULTS

This appendix documents the questions, component construction, interventions, selection rules, and results underlying Sec. 3. The task is always to name the continent of one country. The tables provide the complete country pools, and the three wordings below determine the full question text for each country. Pairing is an analysis construction: each country is submitted in a separate model input.

B.1 QUESTION WORDING AND INPUT CONSTRUCTION

For a country s, the raw question text $p_{\tau,s}$ uses one of the following wordings, followed on a new line by the instruction *Answer with only the continent name.*

Role	Question text
Construction	Which continent is {country} located in?
Evaluation 1	On which continent is {country} located?
Evaluation 2	What continent is {country} in?

The country name replaces {country}. The checkpoint's chat template then adds user and assistant-generation markers, and its tokenizer converts the formatted text to $x_{\tau,s}$. Thus the final input position is determined after formatting and tokenization, not by locating the final printed word *name*. The model and tokenizer identities and rendered-input hashes are retained in the frozen experiment records.

The construction wording supplies one state per fitting country for the content references and supplies the paired states used to construct each request direction. The other two wordings supply the intervention-evaluation questions. There are six content-fitting inputs, 54 direction-construction inputs from the 27 screening and validation pairs, and 108 evaluated questions (12 screening and 96 validation). The complete 168 raw question texts and their roles are included in `evidence/natural_question_inputs/questions.jsonl`; the file contains questions, not answer strings appended as factual context.

B.2 Fitting Countries

The content fitting set contains exactly two countries per continent:

Continent	Country 1	Country 2
Africa	Kenya	Nigeria
Asia	China	Thailand
Europe	France	Spain

For each country, only the construction-wording question contributes a vector to the fitted mean. Consequently, the full mean averages six vectors and each continent mean averages two. No screening or validation country contributes to these content fits.

B.3 Screening and Validation Pairs

The following three pairs select the source and receiving layers. Each pair contributes two countries under both evaluation wordings, giving four evaluated questions:

First country	Continent	Second country	Continent
Egypt	Africa	India	Asia
Vietnam	Asia	Germany	Europe
Portugal	Europe	Morocco	Africa

The 24 disjoint validation pairs below test the fixed choice. No country in this table occurs in the fitting or screening lists:

First country	Continent	Second country	Continent
Angola	Africa	Japan	Asia
Cambodia	Asia	Hungary	Europe
Albania	Europe	Gabon	Africa
Benin	Africa	South Korea	Asia
Malaysia	Asia	Iceland	Europe
Austria	Europe	Malawi	Africa
Botswana	Africa	North Korea	Asia
Philippines	Asia	Ireland	Europe
Belgium	Europe	Mali	Africa
Burkina Faso	Africa	Nepal	Asia
Brunei	Asia	Italy	Europe
Bulgaria	Europe	Mozambique	Africa
Burundi	Africa	Bhutan	Asia
Maldives	Asia	Netherlands	Europe
Croatia	Europe	Namibia	Africa
Cameroon	Africa	Pakistan	Asia
Qatar	Asia	Norway	Europe
Czechia	Europe	Niger	Africa
Chad	Africa	Sri Lanka	Asia
Kuwait	Asia	Poland	Europe
Finland	Europe	Rwanda	Africa
Ethiopia	Africa	Laos	Asia
Oman	Asia	Romania	Europe
Greece	Europe	Senegal	Africa

Each validation pair again supplies four evaluated questions. The pair, with all its wordings and intervention conditions kept together, is the independent resampling unit. The construction and evaluation wordings identify different experimental roles, not additional model inputs supplied alongside a question.

B.4 State Coordinates and Request Construction

The natural-question protocol supplies the intervention definitions omitted from Sec. 3. Qwen-2.5-3B-Instruct is evaluated with fixed weights in float32. At every layer, fitting, recording, and intervention use the decoder-block output at the final position of the rendered chat prompt, immediately after the assistant prefix and before the model predicts the first answer token. The state is recorded before the model's final normalization. The recording position differs from the final recorded point in the original paired-task profiles in Sec. 4; the natural-question protocol uses one coordinate convention throughout.

The six construction-wording fitting questions define the overall mean $\boldsymbol{\mu}^\ell$ and the three continent means $\boldsymbol{\mu}_k^\ell$, each averaging two recorded final-position states. We stack the three centered continent means as rows and compute their singular value decomposition. Let $\boldsymbol{B}^\ell$ contain the two leading right singular vectors as columns (Halko et al., 2011, Section 3.2.2). The fitted continent projector is

$$\boldsymbol{P}^\ell = \boldsymbol{B}^\ell (\boldsymbol{B}^\ell)^\top. \tag{8}$$

The two columns of $\boldsymbol{B}^\ell$ are orthonormal, so $\boldsymbol{P}^\ell$ extracts the part of a hidden-state vector lying along the fitted differences among the three continents. These fitted quantities are shared by all evaluated inputs at the layer.

For one pair of construction questions, let Δ^ℓ be the first question's recorded state minus the second's, and let $\Delta^\ell_{\text{cont}} = \boldsymbol{P}^\ell \Delta^\ell$ be its fitted continent component. We remove $\Delta^\ell_{\text{cont}}$ and normalize the remainder:

$$\boldsymbol{r}^\ell = \texttt{Normalize}\big(\Delta^\ell - \Delta^\ell_{\text{cont}}\big). \tag{9}$$

For a vector whose length exceeds 10^{-12}, `Normalize` rescales the vector to unit length and returns the zero vector otherwise. The zero branch is never used in the evaluated pair–layer constructions: every archived pair–layer group contains a nonzero route edit, and the smallest groupwise maximum actual edit norm is 0.0538. Hence every $\boldsymbol{r}^\ell$ used in the subsequent interventions has unit norm. The direction is constructed separately for each question pair, but never refitted on either evaluation wording.

The archived inputs use one construction wording and two evaluation wordings; Appendix B lists all three. Holding the country pair fixed while changing the wording tests whether the pair-specific direction affects questions beyond the exact construction prompt. The cross-wording test does not make the direction global: a separate direction is still constructed for each country pair. For either evaluation wording, q and $\bar{q}$ denote the two questions about the same countries as the construction pair, and $\boldsymbol{h}_q^\ell$ and $\boldsymbol{h}_{\bar{q}}^\ell$ are their recorded states. The signed coefficient is measured relative to the midpoint of these evaluated questions, not the construction states used to obtain the direction:

$$\alpha_q^\ell = (\boldsymbol{r}^\ell)^\top \left(\boldsymbol{h}_q^\ell - \tfrac{1}{2}(\boldsymbol{h}_q^\ell + \boldsymbol{h}_{\bar{q}}^\ell)\right). \tag{10}$$

The deletion intervention subtracts $\alpha_q^\ell \boldsymbol{r}^\ell$; the replacement intervention subtracts twice that vector. Each operation changes the final prompt-position vector at one intervention block; the input text, weights, and other positions at that block remain unchanged. Subsequent blocks and answer probabilities are recomputed. Each targeted edit has eight separate Gaussian random controls, scaled to its actual written length. Deletion and replacement therefore have their own control lengths.

B.5 Proofs of the Pair-Specific Edit Propositions

Proof of Proposition 1. The construction above gives $\|\boldsymbol{r}^\ell\|_2 = 1$, and α_q^ℓ is the coordinate of evaluation state $\boldsymbol{h}_q^\ell$ relative to the midpoint of the two evaluation states. Therefore,

$$(\boldsymbol{r}^\ell)^\top \left(\boldsymbol{h}_{q,\mathrm{del}}^\ell - \tfrac{1}{2}(\boldsymbol{h}_q^\ell + \boldsymbol{h}_{\bar{q}}^\ell)\right) = \alpha_q^\ell - \alpha_q^\ell = 0.$$

The change $\boldsymbol{h}_{q,\mathrm{del}}^\ell - \boldsymbol{h}_q^\ell$ is parallel to $\boldsymbol{r}^\ell$, so every component orthogonal to $\boldsymbol{r}^\ell$ is unchanged. □

Proof of Proposition 2. Applying the replacement intervention gives

$$(\boldsymbol{r}^\ell)^\top \left(\boldsymbol{h}_{q,\mathrm{rep}}^\ell - \tfrac{1}{2}(\boldsymbol{h}_q^\ell + \boldsymbol{h}_{\bar{q}}^\ell)\right) = \alpha_q^\ell - 2\alpha_q^\ell = -\alpha_q^\ell.$$

Thus the signed coordinate has the same magnitude on the paired side of the midpoint. As above, the edit is parallel to $\boldsymbol{r}^\ell$ and leaves every orthogonal component unchanged. □

B.6 Outcome Scores, Layer Selection, and Uncertainty

Here x denotes a generic evaluated question and plays the same role as q in the main text. Let a_x^+ and a_x^- be the requested and paired continents' answer strings. For an answer $a = (a_1, \ldots, a_{T_a})$, e indexes the unmodified condition, one targeted edit, or one matched random control. We separately compute the answer scores for the requested and paired answers and take their difference:

$$\begin{aligned} s_{x,2}^e &= \frac{1}{T_{a_x^+}} \sum_{u=1}^{T_{a_x^+}} \log p^e(a_{x,u}^+ \mid x, a_{x,<u}^+), \\ s_{\bar{x},2}^e &= \frac{1}{T_{a_x^-}} \sum_{u=1}^{T_{a_x^-}} \log p^e(a_{x,u}^- \mid x, a_{x,<u}^-), \\ s_{2,x}^e &= s_{x,2}^e - s_{\bar{x},2}^e. \end{aligned} \tag{11}$$

The model scores each fixed candidate with its own preceding tokens; this is not unrestricted generation accuracy. A larger component score means that the model assigns a higher mean token log probability to that answer, and a positive difference $s_{2,x}^e$ means that the requested continent is preferred to the paired continent. The archived variable m_x^e is this difference.

At a later layer $\ell' > \ell$, define $\boldsymbol{z}_x^{\ell',e} = \boldsymbol{P}^{\ell'}(\boldsymbol{h}_x^{\ell',e} - \boldsymbol{\mu}^{\ell'})$ and $\boldsymbol{\kappa}_k^{\ell'} = \boldsymbol{P}^{\ell'}(\boldsymbol{\mu}_k^{\ell'} - \boldsymbol{\mu}^{\ell'})$. We likewise compute separate fitted-state scores and their difference:

$$\begin{aligned} s_{x,1}^{\ell',e} &= -\|\boldsymbol{z}_x^{\ell',e} - \boldsymbol{\kappa}_{k_x^+}^{\ell'}\|_2^2, \\ s_{\bar{x},1}^{\ell',e} &= -\|\boldsymbol{z}_x^{\ell',e} - \boldsymbol{\kappa}_{k_x^-}^{\ell'}\|_2^2, \\ s_{1,x}^{\ell',e} &= s_{x,1}^{\ell',e} - s_{\bar{x},1}^{\ell',e}. \end{aligned} \tag{12}$$

Here k_x^+ and k_x^- are the requested and paired continent labels. Each component score is the negative squared distance to its continent reference, so a larger value means that the projected state is closer to that reference. A positive difference $s_{1,x}^{\ell',e}$ means that the projected state is closer to the requested continent reference than to the paired reference. The archived $\gamma_x^{\ell',e}$ denotes this difference. Projectors, centers, and continent references are fixed across intervention conditions.

The main-text table reports the mean component scores under no intervention, deletion, replacement, and the random interventions matched separately to deletion and replacement. Each random row first averages its eight matched random conditions for each question. Fitted-state component scores are displayed in thousands. The resulting values and their source records are retained in `evidence/natural_question_protocol/paired_score_table.csv` and `evidence/natural_question_protocol/paired_score_rows.jsonl`; the original difference scores remain in `evidence/natural_question_protocol/main_table_scores.csv`.

For either outcome, damage is the unmodified score minus the edited score. Within each country pair, we average its four questions separately for the targeted edit and each random control. We then average over country pairs and subtract the largest random-control mean. For the fitted score observed at a later layer, with $\mathcal{X}$ the evaluated inputs and $N = |\mathcal{X}|$, this is

$$\Delta s_{1,c}^{\ell\to\ell'} = \frac{1}{N}\sum_{x\in\mathcal{X}}(s_{1,x}^{\ell',0} - s_{1,x}^{\ell',c}) - \max_{j=1,\dots,8}\frac{1}{N}\sum_{x\in\mathcal{X}}(s_{1,x}^{\ell',0} - s_{1,x}^{\ell',c,j}). \tag{13}$$

Here c identifies the deletion or replacement intervention at the fixed source ℓ, and j indexes its controls. Replacing s_1 by s_2 gives $\Delta s_{2,c}^{\ell}$, the answer-preference effect. Equal pair sizes make the displayed input average equivalent to averaging pair means. The maximum is taken after averaging, not independently for each question. The two scores have different units.

Only the three screening pairs select where to intervene and where to observe the later fitted score. We enumerate intervention layers and strictly later observation layers. A candidate requires positive deletion and replacement answer effects, at least 80% positive unmodified fitted scores at the observation layer, and positive deletion and replacement later-knowledge effects. Among candidates, we first maximize the deletion answer effect to choose the intervention layer, then maximize the deletion later-knowledge effect to choose the observation layer; ties are resolved by choosing earlier layers. This selects $32 \to 36$, frozen before evaluating the 24 validation pairs. All 24 validation pairs pass the baseline gate: for every country request under every recorded wording, the requested continent receives a higher answer score than the paired continent.

Intervals use 2,000 bootstrap samples of 24 country pairs with replacement, keeping each sampled pair's four questions and all conditions together. Each resample recomputes the strongest random-control mean; the 2.5th and 97.5th percentiles give the interval (Efron & Tibshirani, 1986). The base seed is 2026090615. Registered answer-deletion, answer-replacement, knowledge-deletion, and knowledge-replacement comparisons use offsets $900000+i$ for $i = 0, 1, 2, 3$. Layerwise answer comparisons use $1000\ell+i$ for deletion, replacement, and knowledge deletion indexed by $i = 0, 1, 2$. Layerwise knowledge comparisons use $200000 + 1000\ell + 10\ell' + i$ for deletion and replacement indexed by $i = 0, 1$. These are pointwise intervals, not simultaneous confidence bands.

B.7 Registered and Layerwise Results

All 24 validation pairs qualify before intervention. At the frozen $32 \to 36$ comparison, answer effects use mean token log-probability differences and knowledge effects use unscaled squared-distance-score differences; Table 1 displays its raw state scores in thousands. The effects are:

Edit	Answer effect [95% interval]	Knowledge effect [95% interval]
Deletion	2.3963 [1.7023, 2.9530]	1024.06 [−116.48, 2043.71]
Replacement	7.1358 [5.3767, 8.5490]	6087.27 [3268.34, 8553.05]

Deletion's later-knowledge interval spans zero, so the registered conjunction of all four positive tests does not pass. Replacement nevertheless moves both the fitted later state and the answer toward the paired fact. For a fixed ordering of each country pair, the expected sign is $\alpha > 0$ for its first member

and $\alpha < 0$ for its second. The complete 36-layer answer table and all intervention–observation comparisons are retained in the accompanying evidence, including comparisons whose intervals span zero.

Two controls delimit the interpretation. First, the fitted continent-knowledge projection is removed before the pair-specific direction is constructed, so the tested coordinate excludes the measured linear continent component at the edited layer. Second, every deletion and replacement intervention is compared with eight independent Gaussian edits matched to the length of the vector actually written. These controls test the fitted linear component and edit length; they do not exclude every nonlinear or unmeasured representation of knowledge.

B.8 Numerical Checks and Failure Counts

We verify the changes actually written at the intervention position after floating-point rounding. For each targeted deletion or replacement intervention, the relative write error is the Euclidean norm of the actual-minus-intended change divided by the intended change's norm. For each random control, the relative length error is the absolute difference between its actual length and the corresponding targeted change's actual length, divided by the latter. For routing deletion and replacement separately, the content-space residual is the norm of the actual change's projection into the fitted continent space divided by the actual change's norm. Every denominator is lower-bounded by 10^{-12}, as in the recorded analysis.

A check fails when its error or residual ratio exceeds 0.05. The analysis pools all recorded question–layer cases from the primary run, including screening and validation questions. The length-error denominator counts individual random changes for routing deletion and replacement, with eight controls per targeted edit; the write-error denominator counts the corresponding targeted edits. The two routing-residual checks each count the relevant routing edits. The failure fraction is the number of failed checks divided by the number checked in that category; each category must have a failure fraction no greater than 0.05. The independent run is evaluated separately and is not added to these denominators.

The observed failure counts and maximum ratios are:

Check	Failures / checks	Maximum ratio
Random-change actual length	0 / 62,208	6.1467×10^{-7}
Targeted actual write	0 / 7,776	7.2315×10^{-6}
Deletion content-space residual	0 / 3,888	9.0379×10^{-7}
Replacement content-space residual	0 / 3,888	4.7777×10^{-7}

The independent run reproduces the selected layers, every registered pass/fail decision, and all displayed estimates after rounding. One of its three shards used an L40 rather than the A6000 used for the corresponding primary shard: 54/162 baseline and 1,296/3,888 layerwise records differ numerically. Each layerwise record contains all intervention conditions, not one of the 7,776 individual routing writes checked above. The largest baseline-margin difference is 1.30×10^{-4}, and all four registered effect estimates differ by less than 0.003. This supports cross-hardware numerical robustness, not exact equality or a positive deletion knowledge result. Complete checks are archived in `evidence/natural_question_protocol/`.

B.9 Request State Before the Country Appears

This separate task tests a request state that cannot yet contain the current country's answer. Each input asks for either a country's continent or its capital, with the relation word before the country name. Two wordings fit the relation contrast: "Requested attribute: {relation}. Country: {country}. Return only the requested value." and "Retrieve this relation: {relation}. Country: {country}. Return only the value." The held-out wording is "Relation to answer: {relation}. Country entity: {country}. Provide only the corresponding value." Here the relation is *continent* or *capital*. Using Kenya as the fitting reference country, we record each layer's state at the last token of that relation word, average the two fitting states separately for the two relations, subtract the capital mean from the continent mean, and normalize the difference. The midpoint of those means sets the coordinate's zero point.

For each question in the held-out wording, we leave its text and model weights fixed and subtract only the naturally present coordinate along this direction at the relation-word position. We compare its answer-margin decrease with three Gaussian edits matched to the actual write length and with the same targeted change written at the country-word position. The margin is the requested relation's answer score minus the other relation's answer score for the same country; each candidate string receives the mean of its token log probabilities, as in Eq. 11. The decrease is the unmodified margin minus the edited margin. Eight countries are each tested under both relations; this quick experiment separates wordings, not fitting and evaluation countries. The relation-word state precedes the country in the input; for a fixed relation, changing the later country gives a maximum recorded state difference of zero. Thus this edited state cannot carry the current country's specific continent or capital answer.

Country	Continent	Capital
Kenya	Africa	Nairobi
China	Asia	Beijing
France	Europe	Paris
Japan	Asia	Tokyo
Germany	Europe	Berlin
Nigeria	Africa	Abuja
Spain	Europe	Madrid
Thailand	Asia	Bangkok

All 16 unmodified questions prefer the correct answer. The quick scan tests layers 0, 4, 8, 12, 16, 20, 24, 28, and 32; the directional rule requires at least 70% of originally correct questions to move in the predicted direction and the mean target effect to exceed both controls. Layers 12, 16, 20, 24, and 28 pass. We show layer 16 because all 16 questions move in the predicted direction there, not because a country-disjoint set selected it. At that layer, targeted deletion lowers the answer margin by 10.51 on average; the wrong-position and strongest matched-random mean effects are 0.92 and 0.88. The continent and capital questions each show the predicted decrease in all eight cases. This quick experiment establishes a directional answer effect for an earlier request state, not the identity of that state with the pair-specific $\mathbf{r}^{\ell}$ or a causal path through a measured later fact.

C Paired-Question Protocol Details

This appendix collects the common construction and scoring details of the paired-country experiments in Secs. 4 and 5. The natural single-country experiment has its own pair-conditioned construction in Sec. 3. The formulas below specify the paired task, the parameter-routing and fitted-knowledge components used in the main text, and the auxiliary selection contrast retained for exclusion and diagnostic controls.

C.1 Task and Data

A question with two possible knowledge targets. Each input lists two countries from different continents and asks about one of them. For example:

> First country: Kenya
> Second country: China
> Which continent is the **first** country located in?
> Answer with only the continent name.

The paired question changes only *first* to *second*. Its answer changes from Africa to Asia. The input supplies country names and the request, but neither country–continent association. The model must therefore use its stored knowledge to answer. Choosing countries from different continents ensures that changing the request also changes the target knowledge.

One association, three answer formats. We vary how the same knowledge must be expressed: a continent name, its adjective, or an arbitrary code. The corresponding answers in this example are:

Answer format	Request Kenya	Request China
Continent name	Africa	Asia
Adjective	African	Asian
Code	dax	wug

The code input provides the mapping Africa = dax, Asia = wug, Europe = blicket. It does not say which continent Kenya or China belongs to. These formats test whether the layerwise pattern depends on how a country–continent association is expressed; they are not three different knowledge domains.

Fitting and held-out evaluation. The original data contain 24 countries from Africa, Asia, and Europe. We divide them into four country-disjoint partitions. Each partition has six countries arranged into three pairs, giving six questions per answer format when both requests are included. Each continent occurs once in the first position and once in the second position. Thus, the requested position is not tied to one continent.

The *fit partition* supplies the hidden states used to construct the measured directions and knowledge space; this does not train or edit model parameters. The *selection partition* screens their effects. *Validation A* and *validation B* test whether those effects recur on new countries. The separate marker task described below supplies only the selection calibration, not the main knowledge answers.

We also evaluate 48 additional countries that occur in none of these four partitions. They form 24 non-overlapping pairs. Two country orders, two question phrasings, and two requests give eight variants per pair, or 192 questions per model. The components are constructed using only the original fit questions, never the new countries. This tests transfer across countries, wording, and order.

Models and baseline answers. We use Qwen-2.5-3B-Instruct (Qwen Team, 2024), Llama-3.2-3B-Instruct (Dubey et al., 2024), and Gemma-3-4B-Instruct (Gemma Team, 2025). We compare the probabilities of the requested and alternative answers, using the sequence score defined below. A model–format combination enters the original main comparison only if it prefers the correct answer on every question in selection and both validation partitions. Gemma's code format reaches five of six in one partition and is reported as diagnostic. All three models prefer the correct candidate on all 192 expanded questions.

C.2 Complete Pair Lists, Templates, and Randomness

The original country pairs are listed below in input order. Each pair contributes one first-country and one second-country request in each answer format.

Partition	First country	Second country
Fit	Kenya	China
	Thailand	France
	Spain	Nigeria
Screening	Egypt	India
	Vietnam	Germany
	Portugal	Morocco
Validation A	Ghana	Indonesia
	Mongolia	Sweden
	United Kingdom	Tanzania
Validation B	Algeria	Bangladesh
	Singapore	Denmark
	Switzerland	Uganda

The full model-specific task files are supplied under `evidence/paired_question_protocols/`. Files named by model and format contain raw prompts, split assignments, expected answers, rendered-input hashes, token counts, and the final input position. They preserve the checkpoint's actual chat formatting; for Llama, the frozen date inserted into the template is `30 Aug 2026`. These formatting fields are held fixed across conditions.

All main prompts begin with two lines, `First country: {first_country}` and `Second country: {second_country}`. The remaining lines depend on the format:

> Which continent is the {first/second} country located in?
> Answer with only the continent name.
>
> What is the adjectival form of the continent where the {first/second} country is located?
> Answer with only the continent adjective, such as African, Asian, or European.
>
> Continent codes: Africa = dax, Asia = wug, Europe = blicket.
> Which code corresponds to the continent of the {first/second} country?
> Answer with only the code.

The braces denote substitutions, not literal text shown to the model. In the transfer supplement, the code assignment is varied over all six permutations while the rest of the prompt is fixed.

The auxiliary task instead supplies records and markers. For example:

> First record: Kenya has marker dax.
> Second record: China has marker fep.
> Which marker belongs to the requested {first/second} record?
> Answer with only the marker.

All actual record and marker assignments are retained in the model-specific calibration task files; the example does not imply that every pair uses the same markers.

The expanded evaluation uses the 24 country pairs listed in Appendix B.3, rather than the original pairs above. Each pair is presented in both country orders, under both requests and two wordings. The first wording is the continent-name question above; the second is exactly *On which continent is the country listed {first/second} located?* Both retain the two-country listing and *Answer with only the continent name.* The resulting 192 questions are included in `evidence/paired_question_protocols/new_questions.json`. These are not the same prompts as the natural single-country questions or the common-protocol direction comparison.

The original layerwise seeds are 202608310301 for Qwen, 202608310302 for Llama, and 202608310303 for Gemma, with eight random controls per intervention. The expanded evaluation uses seed 2026090509, eight controls, and 2,000 country-pair bootstrap samples. Per-record random streams combine the experiment seed with the integer represented by the first 16 hexadecimal characters of the record identifier's SHA-256, modulo $2^{63} - 1$, before initializing the NumPy generator. Model identities and weight/tokenizer hashes are supplied in `figures/narrow_lifecycle_v2/model_contracts.json`; the expanded contract is included with the task files.

In the expanded evaluation, model forward computation uses bfloat16. Captured decoder-block outputs are converted to float32 and then NumPy float64 for offline fitting calculations. Fitting, reading, and intervention refer to the block output before final normalization; fitting is repeated from the original fitting questions at this location, not by reusing the original post-normalization final-layer fit.

The transfer and fitting-sensitivity supplements reuse the original validation A/B questions and do not add independent countries. The expanded route/knowledge comparison, wrong-label control, and repeated deletion share the expanded 24 pairs. The common-protocol direction comparison also reuses these pairs, but has its own wording subset and float32 model computation. These are complementary measurements, not separate populations whose sample sizes can be added together.

C.3 Measured Components and Auxiliary Calibration

Throughout this appendix and the archived numerical tables, the answer margin m_x^e is the requested-minus-paired answer score s_2 used in Secs. 4–5, and the knowledge score $\gamma_x^{\ell,e}$ is G_q^ℓ under condition e. The index x denotes a complete input, equivalent to the main text's question index q once formatting is fixed; superscript e identifies the intervention condition. Sec. 3 uses analogous scores for a separate natural-question experiment; its scores are not pooled with those here.

The main comparison measures answer-supporting content and parameter routing within the same recorded state. An auxiliary marker contrast is used only to exclude selection of information already supplied in the prompt and to retain the archived diagnostic controls.

Let $\boldsymbol{h}_x^\ell \in \mathbb{R}^H$ be the recorded hidden-state vector for question x at layer ℓ. Here, H is the model's hidden-state dimension and $\ell \in \{1, \ldots, L\}$ indexes its L decoder layers. We use the final input-token position after the model's chat template, immediately before the answer continuation, and call it the *question-end position*. The same position is used across layers. Extraction details, including the final normalization, are given in Appendix D.

Target-knowledge state. For a request about Kenya, the target knowledge is Kenya's continent, Africa; for a request about China, it is Asia. We fit a two-dimensional space from the mean states for Africa, Asia, and Europe using principal-component analysis (Shlens, 2014). Its orthonormal basis is $\boldsymbol{U}^\ell \in \mathbb{R}^{H\times 2}$, and $\boldsymbol{P}^\ell = \boldsymbol{U}^\ell(\boldsymbol{U}^\ell)^\top$ projects a state onto that space. We denote the six main-task fitting questions by $\mathcal{X}_{\text{fit}}$ and their two questions requesting continent k by $\mathcal{X}_{\text{fit},k}$. Their vector means are $\boldsymbol{\mu}^\ell = \frac{1}{6}\sum_{x\in\mathcal{X}_{\text{fit}}} \boldsymbol{h}_x^\ell$ and $\boldsymbol{\mu}_k^\ell = \frac{1}{2}\sum_{x\in\mathcal{X}_{\text{fit},k}} \boldsymbol{h}_x^\ell$. Each model, layer, and answer format has its own fitted space.

Parameter-retrieval routing. We construct a candidate from the mean state for first-country requests minus the mean state for second-country requests. This contrast changes which stored association is requested. We test whether a direction outside the fitted knowledge space influences the answer and the target-knowledge state in later layers. We use *parameter routing* for this measured direction, as in the Introduction.

Auxiliary selection contrast. The auxiliary input supplies arbitrary markers, such as Kenya = dax and China = fep, and asks for the marker of the first or second record. Because the markers are already supplied, its first-minus-second contrast captures selection between prompt contents. The main comparison does not treat this contrast as a third function; it is used to remove this alternative from the paired-task parameter-routing direction. Archived transfer and calibration checks are retained in the later appendices.

To construct paired-task parameter routing, we remove the raw request contrast's knowledge-space projection and its overlap with the knowledge-excluded auxiliary contrast, then normalize the remainder. The superscript global identifies the resulting direction shared by all country pairs within one model and answer format. Appendix D.2 gives the full construction and the archived auxiliary direction.

C.4 Component Deletion and Matched Random Changes

We delete one component while leaving model parameters and all other token positions unchanged. Let $\boldsymbol{b}^\ell$ be the midpoint of the two main-task fit means for first- and second-country requests. The state changes used for the two main components are

$$\begin{aligned} \Delta\boldsymbol{h}_x^{\ell,\mathrm{par}} &= -\big[(\boldsymbol{r}^{\ell,\mathrm{global}})^\top(\boldsymbol{h}_x^\ell - \boldsymbol{b}^\ell)\big]\boldsymbol{r}^{\ell,\mathrm{global}}, \\ \Delta\boldsymbol{h}_x^{\ell,\mathrm{know}} &= -\boldsymbol{P}^\ell(\boldsymbol{h}_x^\ell - \boldsymbol{\mu}^\ell). \end{aligned} \tag{14}$$

The superscripts par and know identify parameter routing and fitted knowledge. We add the chosen vector to the specified decoder-layer output at the question-end position, then continue computation. A standard layerwise test intervenes at one layer only; repeated-deletion checks in Appendix E apply the operation at consecutive layers. Neither protocol intervenes at later answer-token positions.

An arbitrary change can also damage an answer. For each main component $c \in \{\mathrm{par}, \mathrm{know}\}$, we therefore generate eight independent Gaussian vectors $\boldsymbol{\epsilon}_x^{\ell,c,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_H)$, with random-control index $j \in \{1, \ldots, 8\}$. Here, $\boldsymbol{I}_H$ is the H-dimensional identity matrix. We scale each random change to the deletion length:

$$\Delta\boldsymbol{h}_x^{\ell,c,j} = \left\|\Delta\boldsymbol{h}_x^{\ell,c}\right\|_2 \frac{\boldsymbol{\epsilon}_x^{\ell,c,j}}{\left\|\boldsymbol{\epsilon}_x^{\ell,c,j}\right\|_2}. \tag{15}$$

Thus, the comparison changes direction while matching the size of the perturbation. It does not replace the entire hidden state with a random state or copy a component from another question. Appendix E describes numerical-precision matching for repeated interventions.

C.5 Information Present and Effects on Computation

Candidate answers and deletion damage. For each question x, the candidate set $\mathcal{A}_x = \{a_x^+, a_x^-\}$ contains the requested answer and the other country's answer in the required format. Let $a = (a_1, \ldots, a_{T_a})$ be a candidate with T_a tokens. Under intervention condition e, its sequence score $\lambda_{x,a}^e$ and the answer margin m_x^e are

$$\begin{aligned} \lambda_{x,a}^e &= \frac{1}{T_a} \sum_{t=1}^{T_a} \log \texttt{Prob}(a_t \mid x, a_{<t}, e; \theta), \\ m_x^e &= \lambda_{x,a_x^+}^e - \lambda_{x,a_x^-}^e, \qquad \Delta m_x^e = m_x^0 - m_x^e. \end{aligned} \tag{16}$$

Prob is the model's next-token probability under its fixed parameters θ; $a_{<t}$ is that candidate's preceding tokens. Condition $e = 0$ means no intervention. A positive margin means the correct candidate receives the higher score. Δm_x^e is the loss in that margin. This is a controlled two-candidate comparison, not unrestricted generation accuracy.

Answer effect beyond random changes. For $N = |\mathcal{X}|$ questions in an evaluation set $\mathcal{X}$, write $\Delta m_x^{\ell,c}$ for true-deletion damage and $\Delta m_x^{\ell,c,j}$ for damage from random control j, both computed by Equation 16. The answer effect is

$$E_c^{\ell,\text{ans}} = \frac{1}{N} \sum_{x \in \mathcal{X}} \Delta m_x^{\ell,c} - \max_{j=1,\ldots,8} \frac{1}{N} \sum_{x \in \mathcal{X}} \Delta m_x^{\ell,c,j}. \tag{17}$$

The superscript ans denotes an answer effect. A positive value means that deleting the fitted component causes greater average damage than every tested random direction. Negative values mean less damage than the strongest random control, not necessarily no raw deletion damage. We average over questions before taking the maximum over controls.

Target knowledge in later states. Let $\boldsymbol{z}_x^\ell = \boldsymbol{P}^\ell(\boldsymbol{h}_x^\ell - \boldsymbol{\mu}^\ell)$ be the projected, centered state. For each continent k in $\mathcal{K} = \{\text{Africa}, \text{Asia}, \text{Europe}\}$, $\boldsymbol{\kappa}_k^\ell$ is its projected fit-mean reference. For question x, let k_x^+ and k_x^- denote the requested and alternative continents. The knowledge score is

$$\gamma_x^\ell = \left\| \boldsymbol{z}_x^\ell - \boldsymbol{\kappa}_{k_x^-}^\ell \right\|_2^2 - \left\| \boldsymbol{z}_x^\ell - \boldsymbol{\kappa}_{k_x^+}^\ell \right\|_2^2. \tag{18}$$

Positive values mean the state is closer to the requested continent's reference. This score concerns the internal knowledge representation, whereas Equation 16 concerns answer-token probabilities.

After intervening at layer ℓ, we measure knowledge-score losses at each later layer $\ell' > \ell$. Let $\Delta\gamma_x^{\ell \to \ell',c}$ be the unmodified score minus the score after true deletion, and let $\Delta\gamma_x^{\ell \to \ell',c,j}$ be the analogous random-control loss. The later-knowledge effect is

$$E_c^{\ell \to \ell',\text{know}} = \frac{1}{N} \sum_{x \in \mathcal{X}} \Delta\gamma_x^{\ell \to \ell',c} - \max_{j=1,\ldots,8} \frac{1}{N} \sum_{x \in \mathcal{X}} \Delta\gamma_x^{\ell \to \ell',c,j}. \tag{19}$$

This tests whether the intervention changes later target knowledge as well as the final answer.

Correct pointing without intervention. We separately measure whether a component distinguishes the correct target in unmodified states. For routing, the reference $\boldsymbol{b}_x^{\ell,\text{pair}}$ is the midpoint of the two states for the same country pair under opposite requests. Set $\sigma_x = +1$ for a first-country request and -1 for a second-country request. The signed routing values are $\alpha_x^{\ell,\text{par}} = \sigma_x(\boldsymbol{r}^{\ell,\text{global}})^\top(\boldsymbol{h}_x^\ell - \boldsymbol{b}_x^{\ell,\text{pair}})$ and $\alpha_x^{\ell,\text{sel}} = \sigma_x(\boldsymbol{o}^\ell)^\top(\boldsymbol{h}_x^\ell - \boldsymbol{b}_x^{\ell,\text{pair}})$. The correct-pointing fractions are

$$\begin{aligned} F_c^\ell &= \frac{1}{N} \sum_{x \in \mathcal{X}} \texttt{Ind}(\alpha_x^{\ell,c} > 0), \quad c \in \{\text{par}, \text{sel}\}, \\ F_{\text{know}}^\ell &= \frac{1}{N} \sum_{x \in \mathcal{X}} \texttt{Ind}(\gamma_x^\ell > 0). \end{aligned} \tag{20}$$

Ind returns one when its condition holds and zero otherwise. The denominator is the number of questions. Routing pointing is a paired-request diagnostic; knowledge pointing tests proximity to the correct reference. These fractions measure information present, while E^{ans} and E^{know} measure the consequences of intervention.

C.6 Proofs of the Fitted-Knowledge Edit Lemmas

Proof of Lemma 1. Because $\boldsymbol{P}^{\ell}$ is an orthogonal projector and $\boldsymbol{z}_q^{\ell} = \boldsymbol{P}^{\ell}(\boldsymbol{h}_q^{\ell} - \boldsymbol{\mu}^{\ell})$ lies in its range,

$$\boldsymbol{P}^{\ell}(\boldsymbol{h}_{q,\mathrm{del}}^{\ell} - \boldsymbol{\mu}^{\ell}) = \boldsymbol{z}_q^{\ell} - \boldsymbol{P}^{\ell}\boldsymbol{z}_q^{\ell} = \mathbf{0}.$$

Thus the edit removes exactly the fitted projection while leaving the complementary component of the edited position unchanged. □

Proof of Lemma 2. Both reference vectors lie in the range of $\boldsymbol{P}^{\ell}$. Therefore,

$$\boldsymbol{P}^{\ell}(\boldsymbol{h}_{q,\mathrm{rep}}^{\ell} - \boldsymbol{\mu}^{\ell}) = \boldsymbol{z}_q^{\ell} + \boldsymbol{\kappa}_{k_q^-}^{\ell} - \boldsymbol{\kappa}_{k_q^+}^{\ell}.$$

The edit therefore replaces the fitted reference by adding the paired-minus-requested difference and changes no component orthogonal to the fitted space. □

C.7 Layerwise Comparisons and Uncertainty

In the original evaluation, each partition is scored separately. A *stable* main-text effect is positive in screening and both validation partitions. Curves average the two validation effects and shade the range between them. This range is not a confidence interval, and a non-stable value is not set to zero. The archived auxiliary-transfer tests additionally apply their marker-task calibration.

For the expanded evaluation, earlier and later layer sets are fixed separately for each model before testing the new countries. We average damages within each country pair and layer set before applying the random correction. We report 95% percentile bootstrap intervals from 2,000 resamples of the 24 country pairs (Efron & Tibshirani, 1986). All variants and conditions of a pair stay together. The same resamples estimate differences between components or layer sets. Appendix E specifies the aggregation and fixed sets. Fresh model-loading repeats check reproducibility; they do not increase the number of independent samples.

D Narrow-Routing Measurement Details

This appendix makes the paired-task measurements reproducible by fixing their state coordinates, fitting sets, calibration rules, and uncertainty interpretation. The recorded-state location exception is distinguished from the intervention location, and the later subsections give the component formulas and controls.

D.1 Model and Task Records

The frozen runs use the released instruction-tuned checkpoints named in Sec. C: Qwen with 36 decoder layers, Llama with 28, and Gemma with 34. The archived contracts record the exact configuration, tokenizer, weight-index, and weight-shard hashes. Copies of these contracts accompany the figure sources. Computation uses Brain Floating Point 16 (bfloat16), a 16-bit floating-point representation, and all parameters remain fixed during the interventions.

One frozen continent-name prompt is:

```
First country:  Kenya
Second country:  China
Which continent is the first country located in?
Answer with only the continent name.
```

The alternative request replaces first with second. The adjective format asks for the adjectival form of the requested country's continent. In the code format, the input additionally supplies the mapping Africa = dax, Asia = wug, Europe = blicket, and asks for the requested country's continent code.

The separate calibration prompt supplies two records: Kenya has marker dax and China has marker fep. It asks which marker belongs to the first or second record. Calibration therefore measures selection of supplied information, while the main task requires the country–continent association from the model.

Each candidate string is appended to the rendered prompt and scored token by token using its own preceding tokens. Intervention is applied at the question-end position of the specified layer. Layer-wise knowledge scores are read at that same position after subsequent layers.

Model	Format	Stable layers			Validation peaks
		Parameter	Auxiliary	Knowledge	(same order)
Qwen	noun	27–29, 31	—	27–36	0.55 / 0.06 / 7.64
Qwen	adjective	30–31	28–30, 32	27–36	0.28 / 0.11 / 6.25
Qwen	code	27–33	—	27–36	1.88 / 0.03 / 9.30
Llama	noun	15, 21, 24	—	14–28	0.10 / 0.05 / 16.16
Llama	adjective	14–18, 20–21, 24	—	13–28	0.18 / 0.03 / 9.75
Llama	code	11–12, 14–26	—	14–28	0.41 / 0.00 / 8.49
Gemma	noun	17, 19–24	—	18–34	1.96 / 0.29 / 19.96
Gemma	adjective	20–24	—	18–34	3.83 / 0.15 / 22.49
Gemma	code*	24–26	—	17–34	0.41 / −0.01 / 10.86

Table 4: Stable 1-based layers and peak answer-margin effects. The three peak values follow the parameter-routing, auxiliary-selection, and target-knowledge column order. The auxiliary contrast is retained as a calibration diagnostic, not as a main-text functional component. Values may be compared within a model–format row, but raw magnitudes are not pooled or ranked across models. "—" means that no layer has a positive effect in the screening split and both validation splits. The peak column still reports the largest validation mean, including non-stable layers. *Gemma code answers fail the behavioral gate in one split and are diagnostic only.

D.2 Fitting and Centering

For each model, layer, and answer format, let $\mathcal{X}_{\text{fit}}$ be its original fitting questions. The fit mean is $\boldsymbol{\mu}^{\ell} = |\mathcal{X}_{\text{fit}}|^{-1} \sum_{x \in \mathcal{X}_{\text{fit}}} \boldsymbol{h}_x^{\ell}$. For each continent $k \in \mathcal{K}$, let $\mathcal{X}_{\text{fit},k}$ contain the fitting questions whose requested country's continent is k. Then $\boldsymbol{\mu}_k^{\ell} = |\mathcal{X}_{\text{fit},k}|^{-1} \sum_{x \in \mathcal{X}_{\text{fit},k}} \boldsymbol{h}_x^{\ell}$. There are six questions in the full fit and two in each continent group, all from the current answer format. We stack the centered continent means into the matrix

$$\boldsymbol{V}^{\ell} = \left[\boldsymbol{\mu}_{\text{Africa}}^{\ell} - \boldsymbol{\mu}^{\ell},\ \boldsymbol{\mu}_{\text{Asia}}^{\ell} - \boldsymbol{\mu}^{\ell},\ \boldsymbol{\mu}_{\text{Europe}}^{\ell} - \boldsymbol{\mu}^{\ell}\right]^{\top} \in \mathbb{R}^{3 \times H}. \tag{21}$$

Its first two right singular vectors form the columns of $\boldsymbol{U}^{\ell}$. Thus $\boldsymbol{P}^{\ell} = \boldsymbol{U}^{\ell}(\boldsymbol{U}^{\ell})^{\top}$ and the continent references are $\boldsymbol{\kappa}_k^{\ell} = \boldsymbol{P}^{\ell}(\boldsymbol{\mu}_k^{\ell} - \boldsymbol{\mu}^{\ell})$. These are separate fits for each answer format, not a shared coordinate system across formats.

For request-conditioned means, superscripts main and cal indicate the main continent task and auxiliary calibration task, respectively; subscripts first and second identify the requested position. All means below use the corresponding task's fit partition. The raw directions are

$$\begin{aligned} \boldsymbol{r}^{\ell,\text{raw}} &= \boldsymbol{\mu}_{\text{first}}^{\ell,\text{main}} - \boldsymbol{\mu}_{\text{second}}^{\ell,\text{main}}, \\ \boldsymbol{o}^{\ell,\text{raw}} &= \boldsymbol{\mu}_{\text{first}}^{\ell,\text{cal}} - \boldsymbol{\mu}_{\text{second}}^{\ell,\text{cal}}. \end{aligned} \tag{22}$$

First remove the knowledge projection: $\widetilde{\boldsymbol{r}}^{\ell} = (\boldsymbol{I}_H - \boldsymbol{P}^{\ell})\boldsymbol{r}^{\ell,\text{raw}}$ and $\widetilde{\boldsymbol{o}}^{\ell} = (\boldsymbol{I}_H - \boldsymbol{P}^{\ell})\boldsymbol{o}^{\ell,\text{raw}}$. The tildes mark these knowledge-excluded candidates. For a vector $\boldsymbol{v} \in \mathbb{R}^H$, define $\texttt{Unit}(\boldsymbol{v}) = \boldsymbol{v}/\|\boldsymbol{v}\|_2$ when its norm exceeds 10^{-12}, and the zero vector otherwise. Remove overlap with the other candidate and normalize:

$$\begin{aligned} \boldsymbol{r}^{\ell,\text{global}} &= \texttt{Unit}\big(\widetilde{\boldsymbol{r}}^{\ell} - \big[\texttt{Unit}(\widetilde{\boldsymbol{o}}^{\ell})^{\top}\widetilde{\boldsymbol{r}}^{\ell}\big]\texttt{Unit}(\widetilde{\boldsymbol{o}}^{\ell})\big), \\ \boldsymbol{o}^{\ell} &= \texttt{Unit}\big(\widetilde{\boldsymbol{o}}^{\ell} - \big[\texttt{Unit}(\widetilde{\boldsymbol{r}}^{\ell})^{\top}\widetilde{\boldsymbol{o}}^{\ell}\big]\texttt{Unit}(\widetilde{\boldsymbol{r}}^{\ell})\big). \end{aligned} \tag{23}$$

Each residual is orthogonal to the other candidate *before that other candidate is residualized*. The two final unit directions therefore need not be mutually orthogonal, and deleting one does not leave

the fitted coefficient of the other unchanged. We use them as two separately defined measurement directions, not as an additive orthogonal decomposition of the hidden state.

Routing deletion uses the fit-derived center $\boldsymbol{b}^\ell = \frac{1}{2}(\boldsymbol{\mu}^{\ell,\text{main}}_{\text{first}} + \boldsymbol{\mu}^{\ell,\text{main}}_{\text{second}})$; knowledge deletion uses $\boldsymbol{\mu}^\ell$, as specified in Equation 14. The balanced requests make these centers equal in exact arithmetic. They are never chosen separately for validation pairs.

For natural routing pointing, let $\bar{x}$ be the question with the same country order and answer format but the opposite request. Its center is $\boldsymbol{b}^{\ell,\text{pair}}_x = \frac{1}{2}(\boldsymbol{h}^\ell_x + \boldsymbol{h}^\ell_{\bar{x}})$. This pair-specific center is used only for the sign diagnostic in Equation 20, not for deletion. The diagnostic compares two unmodified requests; it is not a standalone classifier given only one question.

Final-layer extraction. The original profiles fit and read states returned by the model's hidden-state interface. Its final entry is after the final normalization, while intervention is applied to the final decoder-block output before normalization. The expanded evaluation aligns fitting, readout, and intervention at decoder-block outputs before final normalization. Nonfinal extraction locations agree. Consequently, the original and expanded final-layer points use different state coordinates. The expanded full-depth results provide the location-aligned final-layer comparison.

D.3 Auxiliary Calibration and Controls

The auxiliary input explicitly assigns unrelated markers to two records. Shifting the question-end state toward the opposite request should change the preferred marker. A layer passes calibration in a partition if baseline marker accuracy is at least 90%, directional switching succeeds on at least two-thirds of questions, and the best of eight matched random controls switches at most one-third. Stable main-task hidden-state-routing effects additionally require calibration in selection and both validation partitions.

For each main-task deletion, eight standard-normal vectors are independently generated, normalized, and scaled to that deletion's Euclidean length. They are unrestricted random changes, not projections removed from a learned random subspace. Seeds combine a fixed experiment seed with the record identifier. Each control is evaluated on the same question, layer, and position as its corresponding deletion.

Main curves use Equation 17: they average each control across questions and then take the maximum. The archived, more conservative alternative instead takes the maximum within each question:

$$E^{\ell,\text{per-question}}_c = \frac{1}{N}\sum_{x\in\mathcal{X}}\left(\Delta m^{\ell,c}_x - \max_{j=1,\ldots,8}\Delta m^{\ell,c,j}_x\right). \tag{24}$$

The superscript identifies the per-question maximum. This value cannot exceed $E^{\ell,\text{ans}}_c$, and the two are not interchangeable. The analogous distinction applies to knowledge-score losses.

The code format supplies an output mapping, but not the queried country–continent association. Correctness is based on the two candidate strings' length-averaged log probabilities, not on unrestricted generation accuracy.

D.4 Replication and Uncertainty

Each evaluated partition has six questions from three country pairs. The two requests for one pair are related observations, not six independent country pairs. The fit, selection, and two validation partitions have disjoint countries. Each run is repeated after reloading the model in a fresh process. The archived comparisons report 5,292 paired component records and 15,876 paired downstream-knowledge records, with no scientific-field mismatches. Input rendering, token positions, and file hashes are retained for reproducibility.

Figures use only the primary runs; reloaded copies are replication checks, not additional samples. Lines average the two validation-partition values at each layer, and shaded ranges show their minimum and maximum. No bootstrap confidence intervals or statistical significance claims are attached to these ranges. Stable layers satisfy the cross-partition positivity criterion of Sec. C. A non-stable layer is not evidence of exactly zero effect.

The unique routing effects exclude directions shared with the other fitted components. The three deletion effects need not add to the effect of deleting their union because subsequent computation is nonlinear. They are reported as changes in answer margin, not fractions of a total control budget.

Figs. 4, 6, and 7 give all three components together for each answer format. Fig. 5 reports the original parameter-routing effects on later knowledge.

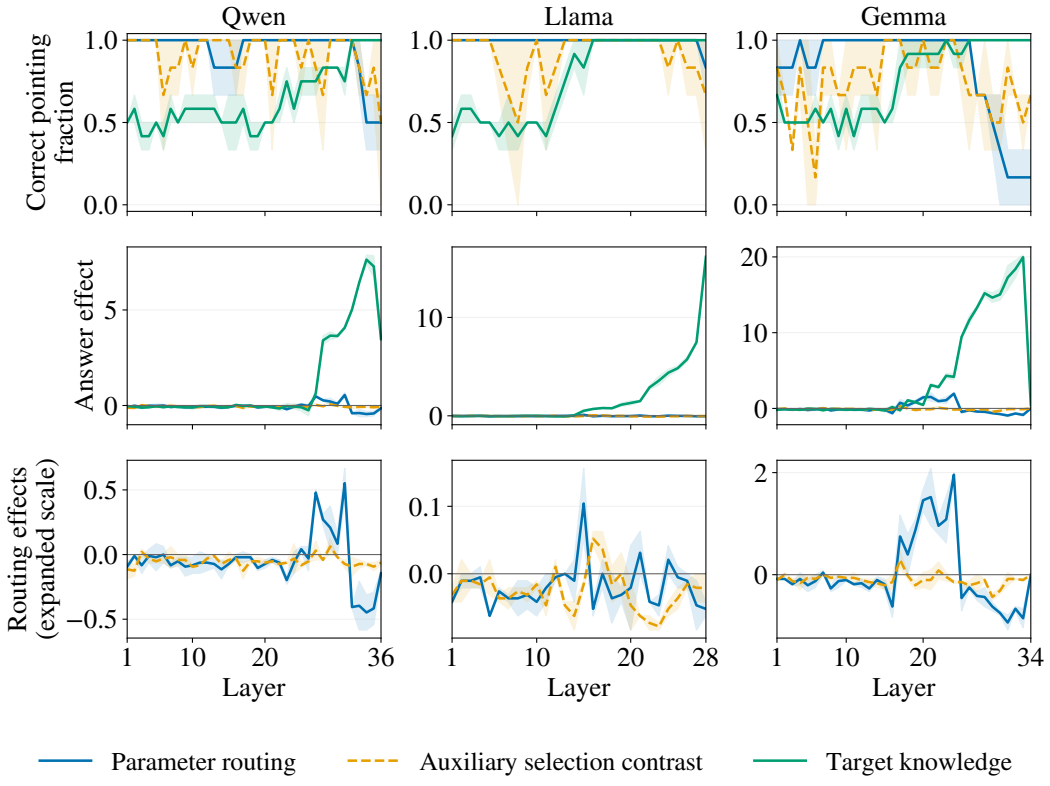


Figure 4: **Three-component profiles on continent-name questions.** The auxiliary selection contrast is a calibration diagnostic, not a third main functional component. Top: correct-pointing fractions. Middle: random-corrected deletion effects on the answer. Bottom: the same routing effects with an expanded vertical scale, without normalization. Lines average validation A and B; shading spans their two values. Orange dashed selection curves include unqualified layers. The presence of a curve does not assert a stable effect.

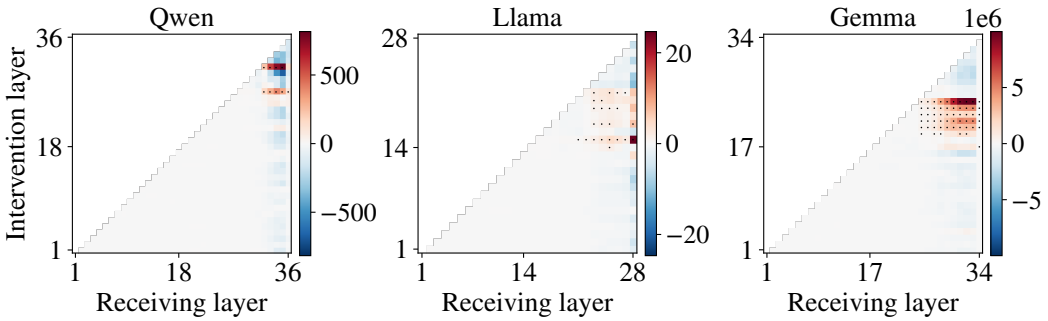


Figure 5: **Paired-task parameter-routing deletion affects later knowledge.** Columns show the continent-name task in each model. Color is the mean random-corrected knowledge loss across the two validation partitions. Black dots mark layer pairs positive in selection and both validation partitions, not statistical significance. Blank cells have no later receiving layer. Model color scales differ because raw knowledge scores are not directly comparable.

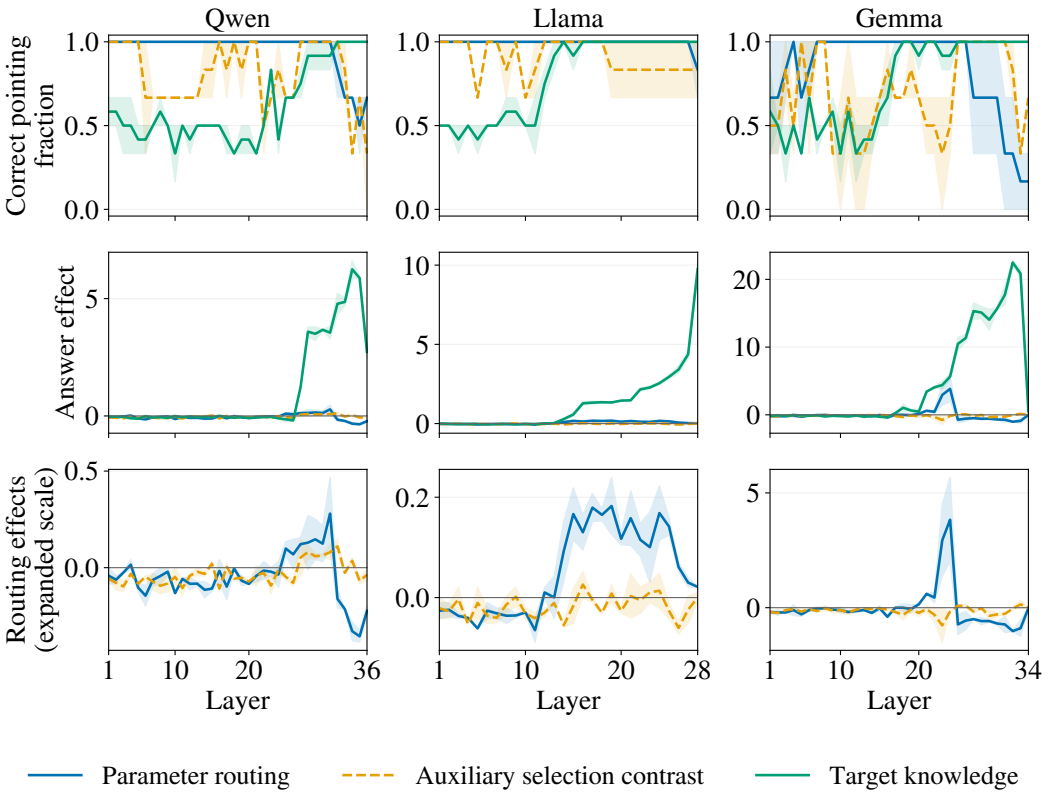


Figure 6: **Adjective-format lifecycle.** Layout and measurements follow Fig. 4. Lines average the two validation results; shading is their range, not a confidence interval. The bottom row expands the routing scale without normalization. Dashed orange lines show the transferred selection candidate, including unqualified layers. Stable effects require the additional criteria in Sec. C; readable or positive values alone do not qualify a layer.

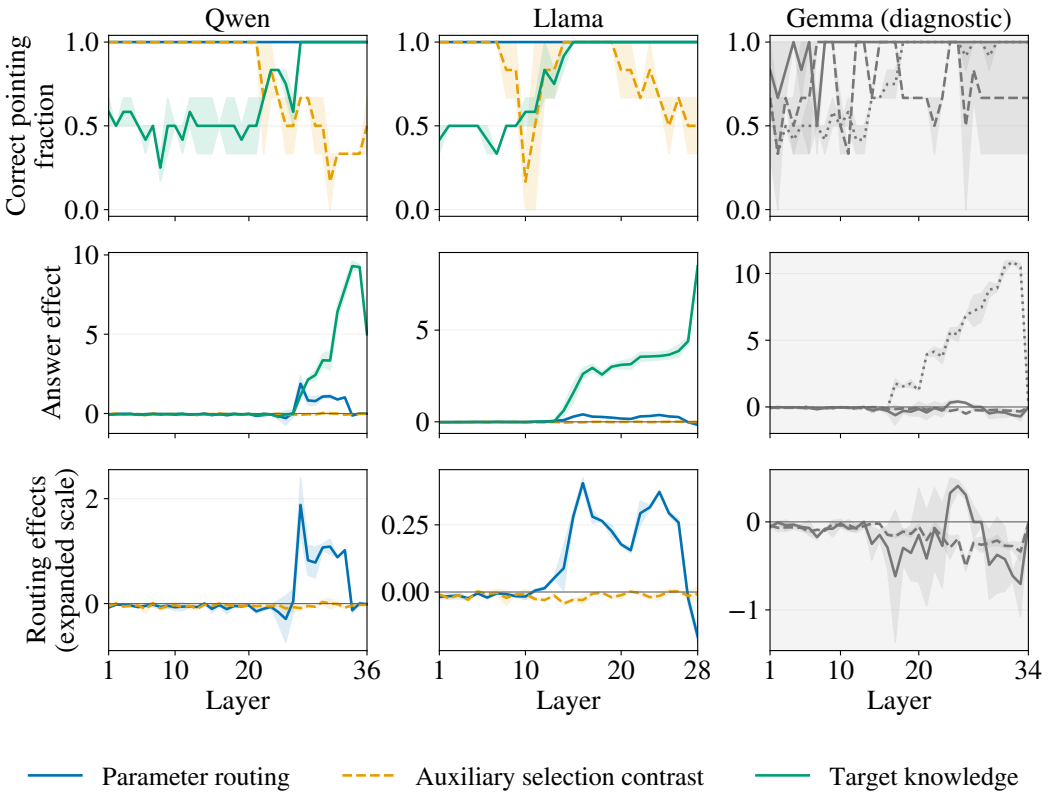


Figure 7: **Code-format lifecycle.** Layout and measurements follow Fig. 4. Gemma's gray column is diagnostic because this model–format combination failed behavioral qualification. Within that column, solid, dashed, and dotted lines denote parameter routing, the auxiliary selection contrast, and target knowledge, respectively. Shading is the range of the two validation results, not a confidence interval. Diagnostic values are retained rather than treated as confirming the main conclusions.

E Expanded Evaluation and Repeated Interventions

This appendix validates the earlier-to-later comparison on new countries and tests alternative explanations through fixed-layer comparisons, wrong-label directions, and repeated deletions.

E.1 New Countries and Fixed Layer Comparisons

The expanded evaluation tests transfer to 48 countries excluded from the original fitting and evaluation partitions. They form 24 non-overlapping country pairs, each with countries from different continents. Swapping country order, rephrasing the question, and requesting either country give eight variants per pair. All 192 questions prefer the correct candidate in every model. We construct the continent-name knowledge space and both routing directions from the original fitting questions using the same formulas; no expanded question is used for fitting.

The earlier and later layer sets were fixed from the original evidence before expanded evaluation. Qwen uses layers 29 and 31 versus 32 and 34; Gemma uses layer 24 versus layers 25 and 30; Llama uses layers 15 and 24 versus layer 28. These are comparison sets, including nonadjacent layers, rather than assertions of continuous processing stages. Expanded fitting, readout, and interventions all use decoder-block outputs before the final normalization. The final-layer measurement therefore refers to the block output, not the normalized state returned by some model interfaces.

The independent resampling unit is a country pair, not one of its eight question variants. Let $\mathcal{Q}$ contain the 24 country pairs and $\mathcal{X}_q$ the eight questions for pair $q \in \mathcal{Q}$. For a selected layer set $\mathcal{L}$, define the pair-averaged true and random damages by

$$\begin{aligned} \overline{\Delta m}^{c}_{q,\mathcal{L}} &= \frac{1}{8|\mathcal{L}|} \sum_{x \in \mathcal{X}_q} \sum_{\ell \in \mathcal{L}} \Delta m_x^{\ell,c}, \\ \overline{\Delta m}^{c,j}_{q,\mathcal{L}} &= \frac{1}{8|\mathcal{L}|} \sum_{x \in \mathcal{X}_q} \sum_{\ell \in \mathcal{L}} \Delta m_x^{\ell,c,j}. \end{aligned} \tag{25}$$

The bars denote averages over that pair's variants and selected layers. The pooled answer effect used in Equation 27 is

$$E^{\mathrm{ans}}_{\mathcal{L},c} = \frac{1}{24} \sum_{q \in \mathcal{Q}} \overline{\Delta m}^{c}_{q,\mathcal{L}} - \max_{j=1,\ldots,8} \frac{1}{24} \sum_{q \in \mathcal{Q}} \overline{\Delta m}^{c,j}_{q,\mathcal{L}}. \tag{26}$$

Here ans denotes the answer outcome, while component labels par and know denote the global route and fitted knowledge. The two comparisons in Sec. 5 are

$$\begin{aligned} \Delta E_{\mathrm{early-late}} &= E^{\mathrm{ans}}_{\mathcal{L}_{\mathrm{early}},\mathrm{par}} - E^{\mathrm{ans}}_{\mathcal{L}_{\mathrm{late}},\mathrm{par}}, \\ \Delta E_{\mathrm{knowledge-route}} &= E^{\mathrm{ans}}_{\mathcal{L}_{\mathrm{late}},\mathrm{know}} - E^{\mathrm{ans}}_{\mathcal{L}_{\mathrm{late}},\mathrm{par}}. \end{aligned} \tag{27}$$

The first measures the decrease in routing dependence; the second compares late knowledge and routing dependence.

For each of 2,000 bootstrap resamples, we sample 24 pairs with replacement and repeat this entire aggregation, including the maximum over controls. All eight variants and all compared conditions of a sampled pair stay together. Using the same resampled pairs for both terms gives the interval for each contrast in Equation 27. These are percentile intervals for the specified comparisons, not simultaneous confidence bands over all layers.

Table 5 gives the exact estimates. Fig. 9 shows the full expanded profiles. Later-layer pointing averages the paired sign diagnostic over the fixed later layer set; it measures request separation and is not an independent classifier for an isolated question.

Each complete experiment is repeated after fresh model loading. Primary and repeated answer margins agree exactly for all 6,912 Qwen, 6,528 Gemma, and 5,376 Llama question–layer records, including every intervention condition. Repeated runs are reproducibility checks and are not pooled as additional observations.

Comparison	Qwen	Gemma	Llama
Earlier routing	0.515 $[0.403, 0.577]$	3.280 $[2.853, 3.681]$	0.102 $[0.075, 0.120]$
Later routing	-0.155 $[-0.215, -0.113]$	-0.148 $[-0.299, -0.057]$	-0.010 $[-0.036, 0.010]$
Routing decline	0.669 $[0.560, 0.758]$	3.428 $[3.031, 3.846]$	0.112 $[0.075, 0.144]$
Later knowledge	6.035 $[5.583, 6.438]$	13.270 $[12.365, 13.889]$	8.281 $[8.053, 8.453]$
Knowledge advantage	6.190 $[5.738, 6.599]$	13.418 $[12.518, 14.065]$	8.291 $[8.066, 8.467]$

Table 5: Expanded answer effects and 95% country-pair bootstrap intervals. Routing decline is the earlier-minus-later routing effect; knowledge advantage is the later knowledge effect minus the later routing effect. Each model uses the same 24 new country pairs and its fixed layer sets. Negative corrected effects mean less damage than the strongest tested random control, not exactly zero raw deletion damage.

E.2 Balanced Wrong-Label Directions

This control tests whether Qwen's earlier routing effect depends on the true request labels. The six original fitting examples admit ten distinct balanced three-versus-three divisions when label reversal is treated as equivalent. One is the true first-versus-second division; the remaining nine define wrong-label controls. For every division we recompute the request-mean difference, remove its fitted knowledge projection and its overlap with the knowledge-excluded auxiliary selection direction, and normalize the result.

Natural-size deletion removes each wrong direction's coefficient from the centered state. Length-matched deletion rescales that vector to the true routing deletion's Euclidean length before insertion into the model. This is vector-length matching; the post-rounding length search described below is used only for repeated random paths. We compare the true direction with the largest mean damage across all nine wrong-label directions. The maximum is recomputed within each country-pair bootstrap resample.

Both comparisons pass in the fixed earlier set, layers 29 and 31 (Fig. 8a). Examined separately, layer 29 has an interval spanning zero for both comparisons; layer 31 has positive intervals. We retain these single-layer results in the figure data. Two complete runs contain 768 question–layer records each, with identical margins for all wrong-label conditions.

E.3 Repeated Deletion and Random Paths

At the question-end position, we remove the routing coefficient from each selected layer's output along that layer's fitted direction. The coefficient is recomputed from the state produced by all previous interventions. Target-knowledge deletion analogously removes the current knowledge-space projection. We test single deletions at layers 29 or 32 and repeated deletions at every layer from 29–36 or 32–36. No intervention is applied at later answer-token positions.

Eight Gaussian random paths control for accumulated perturbation. At each active layer, a path's perturbation is matched to the actual change on the corresponding true-deletion trajectory. For the repeated routing comparison, we adjust its scale to match the length after bfloat16 rounding, without using answer or knowledge scores to select the scale. The search uses 129 logarithmically spaced scales from 0.01 to 100 times the target length, followed by two local refinements with 33 scales each. For repeated knowledge deletion, the original controls already meet the length criterion.

We require relative length error at most 5% at at least 95% of the active positions. After rounding-aware matching, only 12 of 12,288 earlier-routing positions and 5 of 7,680 later-routing positions exceed the error threshold. The original routing controls exceeded it at about 8–9% of positions; Figs. 8b,c use the corrected controls. We also check the component remaining after deletion. Its length is divided by the largest of the incoming length, one tenth of its mean fitting-example length, and a numerical floor of 10^{-12}. We require this ratio to be at most 5% at at least 95% of active positions. The target deletions satisfy this criterion.

To test whether repeated later routing deletion adds a practically meaningful loss, let $\overline{\Delta m}_q^{\text{repeat}}$ and $\overline{\Delta m}_q^{\text{single}}$ be the raw damages averaged over pair q's eight variants for deletion at layers 32–36 and

at layer 32 alone, respectively. Let $\overline{m}_q^0$ be that pair's mean unmodified answer margin. The relative added loss is

$$\rho = \frac{1}{24} \sum_{q \in \mathcal{Q}} \frac{\overline{\Delta m}_q^{\text{repeat}} - \overline{\Delta m}_q^{\text{single}}}{\overline{m}_q^0}. \tag{28}$$

The denominator is positive because all baseline questions pass qualification. This quantity uses raw deletion damage, not the random-corrected effect. Its one-sided 95% bootstrap upper bound is compared with the 10% bound fixed before this experiment. Repeated target-knowledge and routing effects use the same pair resamples when their difference is estimated.

For Qwen, ρ is 0.21%, with a one-sided 95% upper bound of 0.45%, below the prespecified 10% threshold. Thus repeated deletion does not reveal a substantial additional late dependence on the global request component, while repeated knowledge deletion remains consequential (Fig. 8b,c).

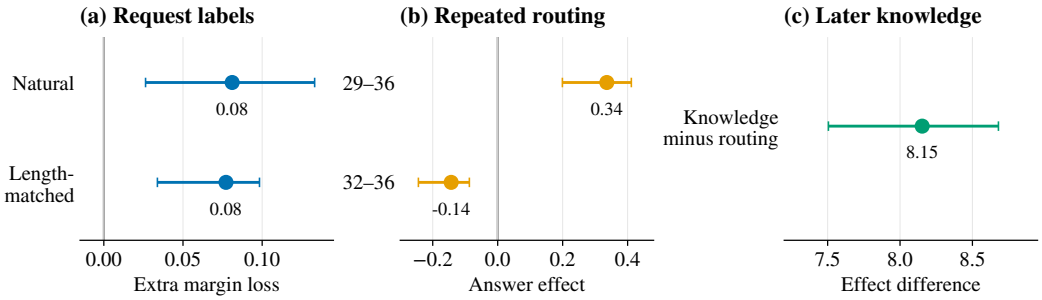


Figure 8: Qwen controls support request specificity and persistent late knowledge dependence. (a) True versus balanced wrong-label directions, pooled over layers 29 and 31. (b) Repeated route deletion with random paths matched to actual edit lengths. (c) Repeated late knowledge deletion versus route deletion. Points and 95% intervals use the same 24 country pairs.

E.4 Additional Knowledge-Relation Screening

A separate extension considered works' original languages, with English, Tamil, and French labels checked against public records. It stopped before component fitting because the frozen candidate pool did not yield disjoint pairs spanning all three language combinations with every question variant correct. Across 768 screening questions, candidate accuracy was 52.3%, with large differences among language labels. Consequently, this extension supplies no intervention evidence about the handoff. The expanded mechanism results above remain about country–continent knowledge.

F Cross-Format and Fitting-Robustness Checks

This appendix tests whether fitted knowledge changes transfer across answer forms and whether the main routing and knowledge effects remain stable when the fitting pairs and random seeds change.

F.1 Fixed Knowledge Changes across Answer Forms

Let k_q^+ and k_q^- denote the requested and alternative continents. Using the noun-derived reference vectors from Equation 5, we apply

$$\Delta \mathbf{h}_q^{\ell,\text{transfer}} = \boldsymbol{\kappa}_{k_q^-}^{\ell,\text{name}} - \boldsymbol{\kappa}_{k_q^+}^{\ell,\text{name}}. \tag{29}$$

The superscript name identifies references fitted only on continent-name questions. The edit is added to the final input-token state at the selected decoder-block output, with model weights and other positions unchanged.

We apply this change to noun, adjective, or code questions at fixed layers: 34 for Qwen, 28 for Llama, and 33 for Gemma. These layers are inherited from the earlier fixed late-knowledge positions

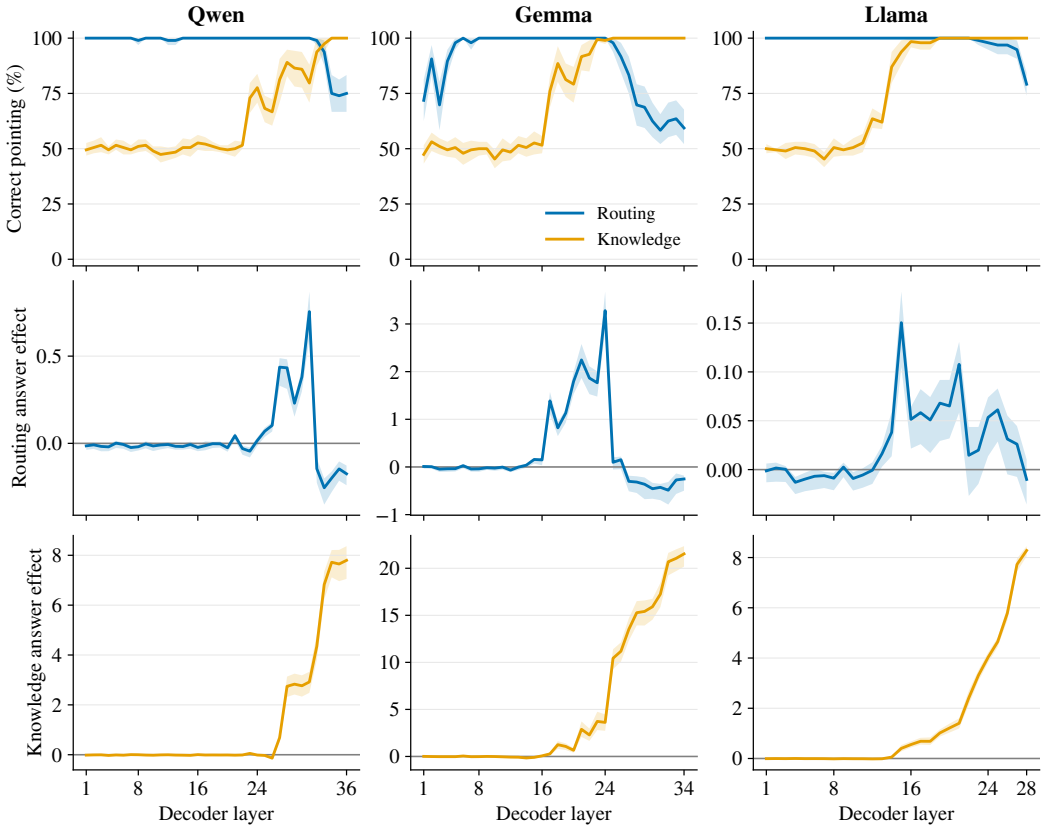


Figure 9: **Expanded depth profiles on 24 new country pairs.** Top: paired routing correct-pointing fractions and target-knowledge correct-pointing fractions. Middle: routing answer effects. Bottom: target-knowledge answer effects. Shading gives pointwise 95% country-pair bootstrap intervals; it differs from the two-partition ranges in Fig. 4. Each column uses its model's full decoder depth and separately fitted components. Routing and knowledge effects have separate vertical scales.

and frozen before this supplement produces new results; no peak is reselected from the transfer-validation results. The noun direction is never fitted again on an adjective or code response.

For descriptive timing only, the archived results also apply paired-reference replacement, using references fitted on continent-name questions, at every decoder layer. Requiring a positive random-corrected mean in both validation groups and positive pair means for at least two of three country pairs yields layers 27–36 in Qwen, 14–28 in Llama, and 18–34 in Gemma. These ranges describe the existing-data scan; they are not independently selected on new questions and do not replace the fixed-layer directional claims in Sec. 4. The per-layer group and pair summaries are in `evidence/cross_format_protocol/transfer_layerwise.csv`.

Each model tests all six code mappings. Baseline qualification is checked separately for each mapping; mappings with a wrong baseline answer remain diagnostic. A transfer passes if its random-corrected shift is positive in both validation partitions and at least two of the three country pairs in each partition have a positive random-corrected pair mean. The reported shift first averages the six questions in each partition and then averages the two partition results. It is not a percentage-point change in accuracy. Both natural answer forms pass in all three models. Code transfer passes for 1/5 qualified mappings in Qwen, 0/4 in Llama, and 0/2 in Gemma.

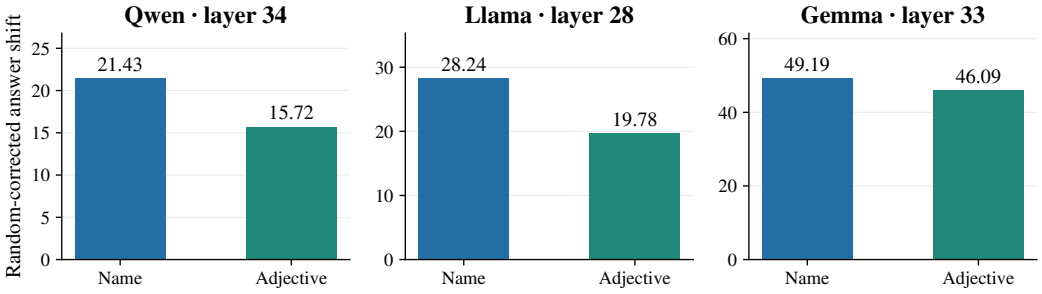


Figure 10: A noun-derived knowledge change transfers to adjective answers but generally not to arbitrary codes. Bars average the random-corrected answer-score shifts of two validation groups; they are not accuracies or confidence intervals. The model-specific magnitudes are not ranked across models. Among baseline-qualified code maps, the transfer criterion passes in 1/5, 0/4, and 0/2 cases for Qwen, Llama, and Gemma, respectively.

For each country pair, the random-corrected mean first averages its two requests' answer-score decreases separately for the targeted change and each random control, then subtracts the largest random-control mean. At least two of three such pair effects must be positive in each validation group. This pair-level condition is separate from the positive whole-group mean condition.

The transfer random seed is 2026090406. Fitting-sensitivity controls use seeds 2026090401, 2026090402, and 2026090403. Both supplements reuse the original validation A/B questions; they assess complementary robustness, not a newly sampled blind confirmation. Complete split, code-map, layer, and control summaries are retained in `evidence/cross_format_protocol/transfer_layerwise.csv` and `evidence/cross_format_protocol/robust_layerwise.csv`.

F.2 Fit and Random-Seed Sensitivity

Four fits use either all three original fitting pairs or omit one pair at a time. Each is tested with three random seeds on both validation partitions. A fit passes when its effect is positive in every seed–partition combination. The registered aggregate criterion requires at least three of the four fits.

At the earlier comparison layers, Qwen passes with 3/4 fits at layers 29 and 31, and Gemma passes with 4/4 at layer 24. Their later routing comparisons pass with 0/4 fits, while content passes with 4/4 fits at every listed comparison layer. Llama routing passes with only 2/4 fits at layers 15 and 24, below the same criterion; its content effect passes with 4/4. These checks concern sensitivity to fitting examples and random seeds. The 24-pair evaluation separately addresses sensitivity to the evaluation countries.

F.3 Raw Damage, Random Damage, and Deletion Length

For Qwen, the transition from layer 31 to 32 changes mean route-deletion damage from 0.8021 to 0.0104 and mean strongest-random damage from 0.2500 to 0.4167. Meanwhile, deletion length increases from 11.2052 to 12.0985. For Gemma, the layer-24 to layer-25 transition changes true damage from 3.2083 to 0.2292 and random damage from 1.2500 to 0.6875. These decompositions support a decrease in true deletion damage, rather than attributing the corrected-score decline only to a changing random baseline.

The three-model supplement includes 20,832 intervention records across primary and fresh-load runs. All 10,416 paired answer-margin records agree. The synchronized archive contains 93 files with matching hashes. Fresh-load repeats check computational reproducibility and are not counted as additional independent observations.

G Route Strength and Cross-Model Timing Diagnostics

This appendix places the natural task's state-relative route strength alongside the paired tasks' normalized request coordinate, answer effects, and later-knowledge effects while keeping their different tasks and normalizations explicit.

G.1 Why the Paired Three-Model Records Are Not a Direct Replication

Sec. 3 measures Qwen on natural single-country questions with one route direction fitted separately for each country pair. Its route-strength denominator is the full hidden-state length. The three-model experiment instead asks paired-country questions, uses one global first-request-minus-second-request direction shared by all evaluated pairs, and divides the natural signed coefficient by half of the fit requests' separation along that direction. A normalized paired-task coordinate of one therefore means that the validation state is displaced by roughly one fitted-request half-separation. It is not the fraction of a hidden state occupied by routing and cannot be compared numerically with S^{route} in Sec. 3.

The three-model analysis retains all three answer forms and both validation partitions. Each model–format combination has six validation country pairs and 12 questions. Lines in Fig. 11 average validation A and B; shading spans their two values. A square marks a layer only when both partitions have positive random-corrected effect and at least 80% of their individual questions have positive targeted-minus-random damage. Primary and fresh-load runs agree exactly for 5,292 component records and 15,876 source-to-receiver knowledge records; the repeat is a computational check, not an additional sample.

G.2 Model-Specific Trajectories

Qwen. The global paired-request coordinate enters a higher mid-layer range before fitted-content deletion becomes stably consequential for continent-name answers at layers 28–30 and 32–36. The global-route deletion itself does not pass the strict split-level answer gate at any continent-name layer, nor does it pass the fixed-later-fact gate. Thus the paired protocol complements, but does not reproduce, the natural-question result: the clean route-strength-to-later-fact sequence remains specific to the pair-conditioned Qwen analysis in Sec. 3.

Llama. The normalized global-request coordinate is already high in early layers, remains high through roughly layer 15, and then declines. Route deletion has no sustained stable answer window for noun or adjective answers; the code task has only two isolated passing layers, 15 and 24. Fitted-content deletion becomes stably consequential from layer 16 for noun answers and layer 15 for adjective and code answers. No source passes the fixed layer-28 fact gate. Llama therefore does not show the same clean route-strength-before-route-to-fact trajectory.

Gemma. The global-request coordinate has a broad middle-layer high region. For continent-name answers, fitted-content deletion passes at layer 21 and then continuously from layers 25–33, while global-route deletion passes only at layer 24. This is a partial overlapping analogue: the request

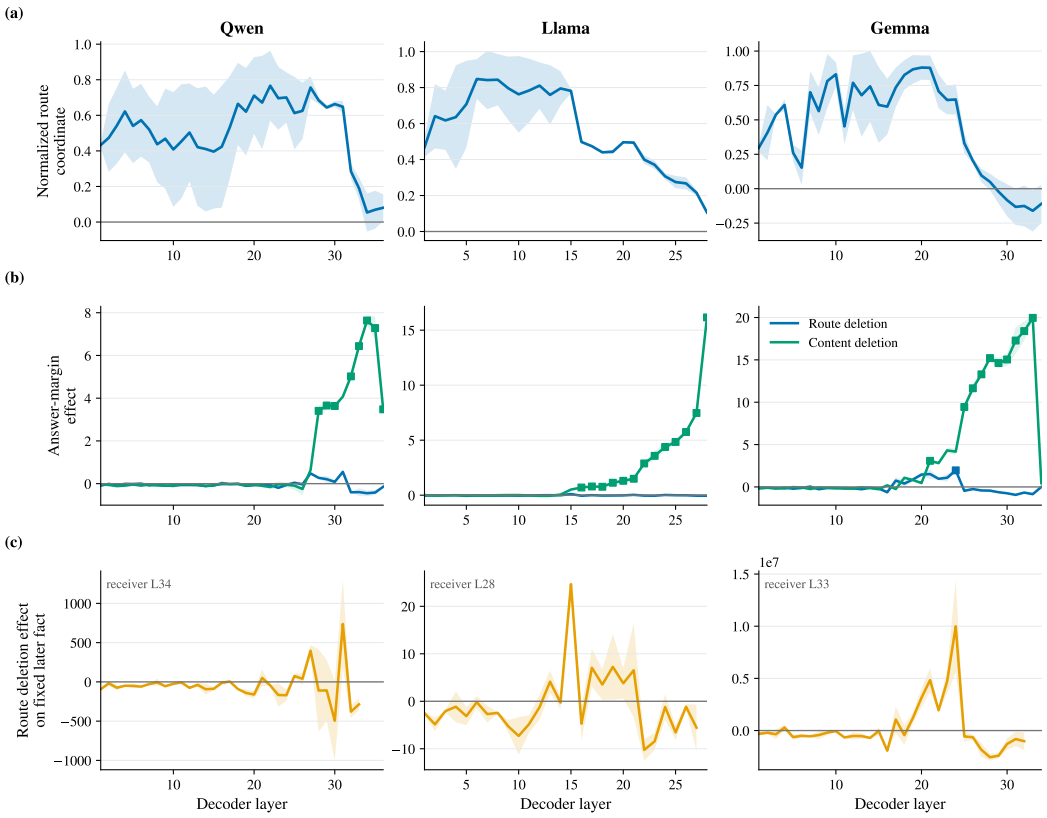


Figure 11: **The paired-country records contain related ingredients but not one cross-model route-first schedule.** Columns show Qwen, Llama, and Gemma for continent-name answers. (a) The unmodified state's signed coordinate along the global request direction, normalized by the fitted request half-separation; values are interpreted within a model and are not whole-state percentages. (b) Answer-margin effects of deleting the global request coordinate or fitted continent content, each relative to eight random changes matched to its actual deletion length. Squares mark layers passing the same split-level stability rule in both validation partitions. (c) The effect of deleting the global route at each source layer on a fixed later fitted-fact score; the receivers are layers 34, 28, and 33 for Qwen, Llama, and Gemma. No continent-name source layer passes the fixed-later-fact stability rule in both partitions. The figure is a diagnostic reanalysis of an existing paired-question protocol, not a direct replication of the Qwen natural-question timing experiment.

coordinate is strong before the local route-answer effect, but content is already becoming consequential and the route does not acquire a sustained fixed-later-fact effect. The adjective and code formats retain sustained content effects without a sustained route-effect window.

Across the three models, the shared result is therefore a distinction among natural coordinate strength, answer dependence, and later-fact dependence, not a universal ordering. Qwen supplies the clearest natural-question route-leading trajectory; Gemma supplies a more overlapping partial analogue in the paired protocol; Llama is a boundary case with dispersed routing effects and earlier sustained content dependence. The plotted values and source hashes are included under `evidence/route_strength_timing/`.

H Common-Protocol Direction Checks

The common-protocol comparison asks which part of the late effect changes when the input task is held fixed and the request measurement changes. We first define the direction and center choices,

then report the late answer and later-knowledge contrasts under matched controls, and finally test how additional random controls affect the natural-question knowledge comparison.

H.1 Direction Construction and Centering

This comparison tests whether a late-layer contrast requires a task change. We hold Qwen-2.5-3B-Instruct weights, float32 computation, final-input decoder-block outputs, answer scoring, and the 24 validation country pairs fixed. The paired templates are "Which continent is the first country located in?", "On which continent is the first country located?", and "What continent is the first country in?", following the same two-country listing and continent-name instruction. Replacing first with second gives the opposite request. The first wording constructs pair-conditioned directions; the other two evaluate them. The natural templates are those in Sec. 3.

Let g index a country pair. Let $x_{\tau,g,i}$ be the complete paired-country input for pair g, template $\tau \in \{A, B, C\}$, and requested position $i \in \{1, 2\}$. The recorded vectors are $\boldsymbol{h}^{\ell}_{x_{\tau,g,1}}$ and $\boldsymbol{h}^{\ell}_{x_{\tau,g,2}}$. Template and requested position construct the input; they are not extra intervention conditions. For compact notation in this comparison, $\boldsymbol{\mu}^{\ell,\mathrm{first}} \equiv \boldsymbol{\mu}^{\ell,\mathrm{main}}_{\mathrm{first}}$ and $\boldsymbol{\mu}^{\ell,\mathrm{second}} \equiv \boldsymbol{\mu}^{\ell,\mathrm{main}}_{\mathrm{second}}$ are the respective averages of the three first- and three second-request states in the paired-task fitting set. Define a knowledge-and-marker exclusion function,

$$\texttt{Exclude}(\boldsymbol{v}; \boldsymbol{P}^{\ell}, \boldsymbol{\eta}^{\ell}) = (\boldsymbol{I}_H - \boldsymbol{P}^{\ell})\boldsymbol{v} - \boldsymbol{\eta}^{\ell}(\boldsymbol{\eta}^{\ell})^{\top}(\boldsymbol{I}_H - \boldsymbol{P}^{\ell})\boldsymbol{v}, \tag{30}$$

where $\boldsymbol{v}$ is an input state difference, $\boldsymbol{P}^{\ell}$ is the paired-task content projector, and $\boldsymbol{\eta}^{\ell}$ is the auxiliary marker contrast normalized after content exclusion. Both request constructions use the same two excluded objects:

$$\begin{aligned} \boldsymbol{r}^{\ell,\mathrm{global}} &= \texttt{Unit}\big(\texttt{Exclude}\big(\boldsymbol{\mu}^{\ell,\mathrm{first}} - \boldsymbol{\mu}^{\ell,\mathrm{second}}; \boldsymbol{P}^{\ell}, \boldsymbol{\eta}^{\ell}\big)\big), \\ \boldsymbol{r}^{\ell,\mathrm{pair}}_{g} &= \texttt{Unit}\Big(\texttt{Exclude}\Big(\boldsymbol{h}^{\ell}_{x_{A,g,1}} - \boldsymbol{h}^{\ell}_{x_{A,g,2}}; \boldsymbol{P}^{\ell}, \boldsymbol{\eta}^{\ell}\Big)\Big). \end{aligned} \tag{31}$$

The superscripts distinguish the shared fit direction from the country-pair direction. The two centers are

$$\boldsymbol{b}^{\ell,\mathrm{fit}} = \tfrac{1}{2}(\boldsymbol{\mu}^{\ell,\mathrm{first}} + \boldsymbol{\mu}^{\ell,\mathrm{second}}), \qquad \boldsymbol{b}^{\ell,\tau,\mathrm{pair}}_{g} = \tfrac{1}{2}(\boldsymbol{h}^{\ell}_{x_{\tau,g,1}} + \boldsymbol{h}^{\ell}_{x_{\tau,g,2}}). \tag{32}$$

Here $\boldsymbol{b}^{\ell,\mathrm{fit}}$ is the fit center $\boldsymbol{b}^{\ell}$ used in Sec. 5, while $\boldsymbol{b}^{\ell,\tau,\mathrm{pair}}_{g}$ is the same pair-specific center denoted $\boldsymbol{b}^{\ell,\mathrm{pair}}_{x}$ when the evaluated input x identifies that pair and template. Their four combinations define c in Equation 33. The replacement intervention doubles that equation's deletion vector. The natural comparison uses only its own pair-conditioned direction and pair center, with knowledge and auxiliary exclusion fitted in the natural-task coordinates. The auxiliary raw marker data are shared, but the content exclusion is task-specific.

For a direction–center combination c and evaluated question q, let $\mathbf{r}^{\ell,c}_{q}$ be its unit direction and $\mathbf{b}^{\ell,c}_{q}$ its center. The deletion is

$$\Delta\mathbf{h}^{\ell,c}_{q} = -\big[(\mathbf{r}^{\ell,c}_{q})^{\top}(\mathbf{h}^{\ell}_{q} - \mathbf{b}^{\ell,c}_{q})\big]\mathbf{r}^{\ell,c}_{q}. \tag{33}$$

Its length equals the absolute projection coefficient, so changing either construction can change the intervention length as well as orientation. Each condition uses eight controls matched to its actual edit length.

H.2 Aggregation, Uncertainty, and Actual Length

Let $\mathcal{G}$ contain the same 24 validation country pairs denoted by $\mathcal{Q}$ in Appendix E and let $\mathcal{X}_g$ contain the four evaluated questions for pair g. For a fixed layer set $\mathcal{L}$, define the pair-averaged loss and the corresponding random-corrected effect:

$$\begin{aligned} \overline{\Delta m}^{c}_{\mathcal{L},g} &= \frac{1}{|\mathcal{L}|\,|\mathcal{X}_g|}\sum_{\ell\in\mathcal{L}}\sum_{x\in\mathcal{X}_g}\Delta m^{\ell,c}_{x}, \\ E^{\mathrm{ans}}_{\mathcal{L},c} &= \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\overline{\Delta m}^{c}_{\mathcal{L},g} - \max_{j=1,\ldots,8}\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\overline{\Delta m}^{c,j}_{\mathcal{L},g}. \end{aligned} \tag{34}$$

The overbar denotes averaging within a country pair and layer set; superscript c, j denotes random control j matched to intervention c. Thus the maximum is taken after, not before, averaging. For

later knowledge, substitute the source-to-receiver knowledge-score loss from Equation 13. All 24 pairs qualify in both tasks before intervention, so qualified-pair and all-pair estimates coincide.

The earlier layers are fixed at 29 and 31, the later set is $\mathcal{L}_{\text{late}} = \{32, 33, 34, 35, 36\}$, and the source-to-receiver knowledge comparison is fixed at 32 to 36. Each bootstrap sample resamples 24 country pairs with replacement. The same sampled pairs are used across conditions, and the maximum random-control mean is recomputed within each sample before taking a condition difference. Intervals are the 2.5th and 97.5th percentiles of 2,000 samples; they are not simultaneous intervals across comparisons. Independent model reloads reproduce the same scientific fields and are not additional statistical samples.

Table 6 gives the individual late answer effects underlying the contrasts. Pair-conditioned deletion remains positive under either center, and also in natural questions; global deletion produces less damage than its strongest matched random control.

Table 6: Late answer effects depend on request-direction construction. Effects average layers 32–36 and four questions per pair, then subtract the strongest mean effect among eight matched random controls. Intervals resample 24 pairs. All rows use float32; the natural-question row has its own fitted knowledge space.

Task	Direction	Center	Effect [95% interval]
Paired	Global fit	Fit	-0.195 [-0.251, -0.160]
Paired	Global fit	Pair	-0.034 [-0.078, -0.009]
Paired	Pair-conditioned	Fit	3.092 [2.464, 3.624]
Paired	Pair-conditioned	Pair	3.051 [2.438, 3.581]
Natural	Pair-conditioned	Pair	2.401 [1.760, 2.971]

Table 7: Direct contrasts of the common-protocol late answer effects. Each row subtracts the second named condition from the first using shared country-pair bootstrap samples. Direction changes retain the stated center; center changes retain the stated direction.

Comparison	Difference [95% interval]
Pair-conditioned minus global; fit center	3.287 [2.674, 3.821]
Pair-conditioned minus global; pair center	3.085 [2.481, 3.629]
Pair minus fit center; global direction	0.161 [0.117, 0.210]
Pair minus fit center; pair-conditioned direction	-0.041 [-0.082, 0.005]
Natural minus paired; pair direction and center	-0.650 [-0.885, -0.395]

Matching each intervention to its own random controls does not equalize the true interventions. Across the 480 paired-task question–layer observations in the later set, mean actual deletion lengths are 16.38 for the global fit-centered condition and 80.47 for the pair-conditioned fit-centered condition. Hence Table 7 identifies dependence on the deletion definition, including its projected length, not orientation alone. In addition, the global direction's earlier effect under this wording subset has an interval spanning zero. This does not isolate a precision effect relative to the older bfloat16 protocol, which also used different wording and order coverage.

Within the same paired-task content coordinates and fit center, layer-32 deletion produces corrected layer-36 knowledge losses of -82.7 [-330.1, 72.2] for the global direction and 2648.4 [1223.3, 3611.0] for the pair-conditioned direction. Their direct difference is 2731.1 [1382.9, 3691.5]. These scores use Equation 12; they are not percentages or cross-task measures of how much knowledge was retrieved.

H.3 Random-Control Sensitivity of Natural-Question Knowledge Effects

This is a post hoc sensitivity analysis of completed records, not a replacement of the registered natural-question test. For all 96 matched questions, baseline and edited answer margins and raw layer-32-to-36 deletion knowledge losses are identical between the two runs compared here. The runs differ in their random controls. We pool their sixteen random controls, retaining the maximum-after-averaging rule and using a common country-pair bootstrap sequence for comparison. The

"Original eight" row comes from the earlier natural-question run used in this sensitivity check, not the separate run reported in Appendix B.7; their raw deletion losses, and therefore their point estimates, differ.

Table 8: Sensitivity of the natural-question layer-32-to-36 deletion knowledge effect to the random controls. Raw true deletion losses are identical. The common bootstrap sequence can produce slightly different interval endpoints from the original registered report.

Random controls	Knowledge effect [95% interval]
Original eight	1102.5 [−47.6, 2128.2]
Additional eight	1376.1 [184.6, 2292.5]
Pooled sixteen	1102.5 [−50.3, 2127.8]

Pooling the controls retains an inconclusive deletion knowledge effect (Table 8). The pooled replacement effect remains positive at 6270.8 [3261.0, 8747.7]. A favorable set of eight controls therefore does not justify upgrading deletion necessity, although the directional replacement finding remains supported.