How Transformers Reject Wrong Answers: Rotational Dynamics of Factual Constraint Processing

Javier Marín*
Groundlens-dev

Revised version, July 7, 2026 (supersedes arXiv:2603.13259v1)

Abstract

When a decoder-only transformer is forced to process matched correct and incorrect single-token continuations of a factual query, the two pathways through hidden-state space diverge in a specific way: displacement vectors from the query-only representation maintain approximately equal magnitude but rotate apart in direction. The angular separation grows through mid-depth, and late layers resolve the asymmetric outcome—a logit-lens preference that, in the incorrect run, falls far below the naive prior of equal probability, corresponding to the model assigning approximately 11.5× more probability to the incorrect token than to the correct one. We characterize this two-phase pattern—rotational divergence in mid-depth followed by late-layer asymmetric commitment—as the empirical geometric signature of what looks externally like the model "rejecting" a wrong continuation, while remaining explicit that it is an *observational* characterization, not a causal account, and that the incorrect-run trajectory admits a second reading (the model conforming to the token it is forced to carry) that only a random-token control can separate. The pattern is consistent across six decoder-only transformers with measurable factual processing, spanning four architecture families (Llama, Mistral, Gemma, StableLM) from 1B to 13B parameters; a seventh model (Qwen2 1.5B), the fifth family, shows a flat profile under the present extraction protocol that is plausibly a tokenizer-fragmentation artefact rather than a real scale floor, so the question of an emergence threshold is left open. Linear probes recover the correct/incorrect distinction at intermediate depth in all six measurable models, and cross-domain probe transfer is structurally asymmetric—a financial–medical corridor transfers far better than transport pairs—a regularity that replicates across all six architectures. On the two models where single-layer activation patching is cleanly interpretable (LLaMA-2 13B and Mistral 7B), patching from the correct run into the incorrect run does not produce a layer band with consistent recovery of the correct token; a third patched model (StableLM-2 1.6B) recovers uniformly at every layer including the first and above the recovery ceiling, a pattern we diagnose as an artefact of its patching code path and exclude. Under this scoped result the late-layer asymmetry is not localized to a single discrete component in the regime tested. Taken together, the evidence is consistent with a distributed-by-trajectory account of factual constraint processing—geometric structure that emerges cumulatively across many layers rather than from a single localized circuit—and inconsistent with the simplest single-layer localized-recall account. We document this structure with controlled forced-completion probing across seven models, three professional domains, and 300 stratified queries.

1 Introduction

Transformer language models generate text that can be factually incorrect. Such failures fall under the broader heading of factual hallucination, a term whose definition varies considerably across

*javier@groundlens.dev

the literature; we adopt a deliberately narrow operational view of factual correctness, detailed in Section 2. Understanding the internal processing that differentiates correct from incorrect factual continuations is a prerequisite for principled mitigation. A growing body of work has established that hidden representations encode linearly separable features related to truthfulness [Burns et al., 2023, Azaria and Mitchell, 2023, Li et al., 2023, Marks and Tegmark, 2024], that factual associations can be localized to specific layers and components [Meng et al., 2022, Geva et al., 2023, 2021], and that intermediate representations can be decoded into output distributions via the logit lens [nostalgebraist, 2020, Belrose et al., 2023]. These approaches are mainly static: a probe at a particular depth, a feature direction in a single layer, an intervention at a localized component. Less is known about the layerwise dynamics: how the distinction between correct and incorrect factual continuations emerges, evolves, and resolves across the full depth of the network. We address this gap with a controlled experimental instrument and a layerwise geometric analysis. Our method, forced-completion probing, builds queries with exactly one single-token correct continuation and exactly one single-token incorrect continuation, and measures five complementary quantities at every layer for both forced runs: trajectory similarity, displacement geometry, linear probe accuracy, logit-lens commitment, and attention allocation. By forcing the model to process both completions of the same query, we obtain matched pairs that isolate the effect of factual correctness from confounds such as token frequency or syntactic complexity. The protocol is methodologically conservative in the following sense: it observes geometric structure under controlled forced continuations but does not by itself bridge to the geometry of hallucinations produced under unconstrained autoregressive generation. We are explicit about this scope throughout.

Scope and stance. We document the layerwise geometric structure that emerges in this regime, characterize its scale and architecture dependence, and report a null result on single-layer activation patching that bounds the causal account our data can support. The null is scoped: it holds on the two architectures where the patch is cleanly interpretable (LLaMA-2 13B and Mistral 7B), while a third patched model (StableLM-2 1.6B) produces a profile we diagnose as an implementation artefact and exclude (Section 4.6, Appendix D). We are deliberately conservative on interpretation: the data are consistent with a distributed-by-trajectory account in which the geometric structure is the cumulative consequence of many small contributions across the network, and also consistent in principle with a localized mechanism that single-layer patching fails to detect. We do not adjudicate between these accounts; we document the structure, report the null result and the excluded artefact plainly, and frame the question of mechanism as open.

2 Related Work

Probing factual representations and recall mechanisms. A growing line of work has documented that hidden representations encode linearly separable features related to truthfulness [Burns et al., 2023, Azaria and Mitchell, 2023, Li et al., 2023, Marks and Tegmark, 2024], that factual associations can be localized to mid-layer MLP modules [Meng et al., 2022, Geva et al., 2023, 2021, Dai et al., 2022], and that knowledge can be edited at those localized sites. These analyses are predominantly static (a probe at a particular depth, an intervention at a localized component) and focus on simple fact-completion tasks. We complement them with a dynamic, layer-by-layer account of matched correct vs. incorrect forced continuations, and we treat our null result on single-layer patching (Section 4.6) as evidence that the localization properties documented in this line do not transfer cleanly to the forced-continuation regime, on the architectures where our patch is cleanly interpretable.

Representation geometry and intermediate-layer decoding. The linear representation hypothesis [Park et al., 2024] and work on structured representations [Gurnee and Tegmark, 2024, Dar et al., 2023, Hernandez et al., 2024, Ethayarajh, 2019] establish that transformer hidden states carry geometric structure. The logit lens provides a way to decode intermediate representations through the unembedding projection [nostalgebraist, 2020, Belrose et al., 2023, Din et al., 2023]. Our approach combines both: a controlled protocol that decomposes displacement geometry into radial and angular components and that uses the logit lens to define commitment ratios for the model's intermediate preference between correct and incorrect tokens.

Scope within the hallucination literature. A model produces a hallucination when its output is not grounded in the facts it should reflect. The literature slices this differently: Maynez et al. [2020] separates faithfulness from factuality; Ji et al. [2023] and Huang et al. [2023] catalog the broader taxonomy; Tam et al. [2023] measures factual consistency in generated summaries; Liu et al. [2026] reframes the question in terms of the model's internal world-model. Across these framings, factually incorrect token-level continuations are recognised as one form of hallucination. This paper does not measure free-generation hallucination. It measures the hidden-state geometry produced when a model is conditioned to output a specific factually-correct or factually-incorrect single-token continuation. The relevance to the broader hallucination question is direct: the geometric signature we document—rotational divergence of equal-magnitude displacement vectors—appears under conditioning, before the model commits to a token. Whether the same signature appears under free generation, and whether it can be used as an early-warning indicator, is the natural next experiment and is beyond the present scope (Section 6).

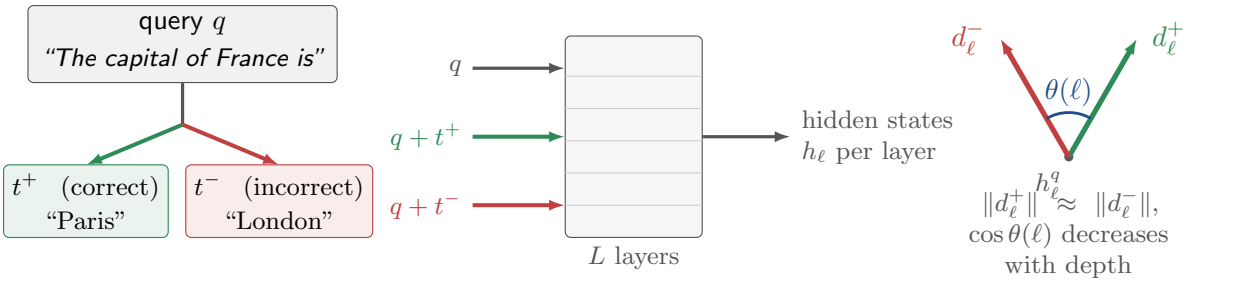
3 Methods

3.1 Forced-Completion Probing

We define forced-completion probing as the following protocol (Figure 1). For each query q we run three forward passes: a *correct run* in which the correct single-token continuation t^+ is appended to the query, an *incorrect run* in which an incorrect but domain-meaningful single-token continuation t^- is appended, and a *query-only pass* with no token appended as a baseline. At every layer ℓ we extract the residual-stream hidden state at the response-token position, yielding $h_\ell^+, h_\ell^- \in \mathbb{R}^d$ from the completion runs and h_ℓ^q from the query-only pass. Hidden states (`float16`, per the model load in Section 3.3) are upcast to `float32` before geometric computation. The protocol's strength is that the matched-pair design isolates the effect of factual correctness from token-frequency, syntactic-complexity and length confounds. Its limit is that a forced continuation is not free generation, and that the incorrect-run trajectory does not by itself distinguish rejection of a wrong fact from accommodation of the token the model is compelled to carry; the random-token control of Section 6 is designed to separate the two.

3.2 Dataset Design

We manually generated 300 queries stratified across three domains (financial compliance, medical protocols, transport regulation) with $n = 100$ each and three category types: deep constraint ($n = 182$, correct answer requires integrating a domain-specific factual rule with the query context where a surface cue would favor the incorrect answer); control ($n = 66$, correct answer is determinate and surface and deep cues agree); and neutral ($n = 52$, the correct/incorrect distinction does not depend on domain-specific factual reasoning, serving as a within-protocol null control following



(a) Matched pair **(b) Three forward runs** **(c) Geometric divergence**

Figure 1: **Forced-completion probing.** A matched single-token pair (a) is processed through three forward passes (b), yielding the layerwise displacement geometry that is the object of study (c). Notation and formal definitions are given in Section 3.1.

Hewitt and Liang [2019]). Each query is human-authored such that (i) the correct t^+ and incorrect t^- continuations both map to single BPE tokens under the studied tokenizers, (ii) both are domain-meaningful rather than arbitrary distractors, and (iii) deep-keyword and surface-keyword spans are character-offset-aligned to the query string for attention analysis. The complete per-domain category mapping, sample queries per category, and full annotation protocol are in Appendix C.

3.3 Models

We evaluate seven decoder-only transformer models across five architecture families: LLaMA-2 13B [Touvron et al., 2023], Mistral 7B v0.3 [Jiang et al., 2023], Llama 3.2 3B [Dubey et al., 2024], Gemma 2 2B [Gemma Team, 2024], StableLM-2 1.6B [Bellagente et al., 2024], Llama 3.2 1B [Dubey et al., 2024], and Qwen2 1.5B [Yang et al., 2024]. Model parameters and layer counts are in Table 1. All models are loaded in `float16` with `attn_implementation="eager"`.

Table 1: Decoder-only transformer models evaluated.

Model	Params	Layers	d
LLaMA-2 13B	13B	40	5120
Mistral 7B v0.3	7B	32	4096
Llama 3.2 3B	3B	28	3072
Gemma 2 2B	2B	26	2304
StableLM-2 1.6B	1.6B	24	2048
Llama 3.2 1B	1B	16	2048
Qwen2 1.5B	1.5B	28	1536

3.4 Measurements

We extract five complementary measurements per query, per layer, per run.

- **(M1) Trajectory similarity**: $\tau(\ell) = \cos(h_\ell^+, h_\ell^-)$ between correct- and incorrect-run hidden states.
- **(M2) Displacement geometry**: displacement vectors $d_\ell^\pm = h_\ell^\pm - h_\ell^q$ decomposed into cosine similarity $\cos(d_\ell^+, d_\ell^-)$ and per-query norm ratio $\eta_i(\ell) = \|d_{i,\ell}^+\|/\|d_{i,\ell}^-\|$ (with mean $\bar{\eta}(\ell)$ across

queries reported in Table 2), separating angular from radial divergence.

- **(M3) Linear probing**: logistic regression with 5-fold stratified cross-validation on hidden states at each layer to predict correct vs. incorrect run.
- **(M4) Logit-lens commitment ratio**: $\kappa(\ell) = \exp(\phi(h_\ell)[t^+])/(\exp(\phi(h_\ell)[t^+])+\exp(\phi(h_\ell)[t^-]))$, where $\phi : \mathbb{R}^d \to \mathbb{R}^{|V|}$ is the unembedding projection, using the normalized logit lens of Belrose et al. [2023]; the naive prior is $\kappa = 0.5$.
- **(M5) Attention allocation**: $\alpha(\ell)$, the fraction of attention from the response-token position directed to manually annotated deep-constraint token spans, averaged across heads.

3.5 Statistical Framework

All pairwise comparisons (deep-constraint vs. neutral queries, correct-run vs. incorrect-run, within-domain vs. cross-domain) use the Mann-Whitney U test with Benjamini-Hochberg false discovery rate correction [Benjamini and Hochberg, 1995] across the layers of each model. Effect sizes are rank-biserial $r = U/(n_1 \cdot n_2)$. Logit-lens commitment asymmetry uses a one-sample t-test against $\kappa = 0.5$ at the minimum-κ layer per model.

3.6 Causal Localization

To test whether a single layer causally produces the layerwise dynamics measured in M1–M5, we run activation patching. The method is the following: for each query, at each layer ℓ, we replace the residual-stream input to layer ℓ in the incorrect run with the corresponding state from the correct run and measure the normalized recovery of the correct-token logit:

$$e(\ell) = (\text{logit}_{\text{patched}} - \text{logit}_{\text{i-baseline}})/(\text{logit}_{\text{c-baseline}} - \text{logit}_{\text{i-baseline}}) \quad (1)$$

$e(\ell) = 1$ indicates full recovery, $e(\ell) = 0$ no causal contribution, and $e(\ell) < 0$ indicates interference (the layer's correct-run state pushes the incorrect-run logit away from recovery). We apply patching on LLaMA-2 13B, Mistral 7B and StableLM-2 1.6B. The alignment between the collected source states and the module receiving the patch is validated for models exposing the `model.model.layers` block container (LLaMA-2, Mistral); StableLM-2 exposes a different container (`transformer.h`) and, as reported in Section 4.6, produces a profile inconsistent with a valid single-layer patch, so we treat it as an artefact and exclude it. Qwen2 1.5B is excluded upstream because its tokenizer fragments the response tokens into multiple sub-tokens for several queries, so the last-token-position extraction reads from an intermediate sub-token rather than from the response prediction. The implementation note on `forward_pre_hook` use under `transformers` ≥ 4.40, the architecture-specific source-alignment issue, and the value-clipping protocol are in Appendix D.

4 Results

4.1 Isometric Rotational Divergence

Table 2 reports the displacement decomposition. Across the six models with measurable factual processing, mean displacement cosine similarity $\cos(d_\ell^+, d_\ell^-)$ drops from 1.00 at the embedding layer to 0.65–0.80 at intermediate depth, while mean norm ratios $\bar{\eta}(\ell)$ remain close to unity throughout (Table 2; quantified below). The model does not distinguish correct from incorrect by making one pathway louder; at the population level the discriminative signal is encoded in the angular

Table 2: **Norm ratio $\bar{\eta}(\ell)$ across models and layer bands.** Cells report mean ± standard deviation of the per-query norm ratio $\eta_i(\ell) = \|d^+_{i,\ell}\|/\|d^-_{i,\ell}\|$, computed over $n = 300$ queries per cell at the layer closest to each target fraction ℓ/L. The mean stays close to unity at every cell across six architectures spanning 1B to 13B parameters, with the largest deviation at 5.5% (Gemma 2 2B, final layer), supporting a population-level isometric divergence (Definition 1). Qwen2 1.5B is reported separately: the bulk of its queries are excluded due to the tokenizer-extraction issue (see text), so the reduced SD reflects sample restriction, not lower spread.

Model	$\ell/L \approx 0.25$	$\ell/L \approx 0.50$	$\ell/L \approx 0.75$	$\ell/L \approx 1.0$
Llama 3.2 1B	1.007 ± 0.045	1.006 ± 0.062	1.003 ± 0.113	1.027 ± 0.111
StableLM 2 1.6B	1.009 ± 0.051	0.996 ± 0.066	0.990 ± 0.099	1.022 ± 0.103
Gemma 2 2B	1.012 ± 0.034	0.996 ± 0.070	0.993 ± 0.115	1.055 ± 0.218
Llama 3.2 3B	1.005 ± 0.059	1.010 ± 0.068	0.998 ± 0.098	1.019 ± 0.077
Mistral 7B	1.004 ± 0.073	0.999 ± 0.082	1.000 ± 0.097	1.014 ± 0.100
LLaMA-2 13B	0.999 ± 0.060	0.992 ± 0.083	0.995 ± 0.108	1.022 ± 0.092
Qwen2 1.5B (excluded)	0.998 ± 0.025	0.998 ± 0.025	0.998 ± 0.025	0.998 ± 0.025

component of the displacement. The mean across queries, $\bar{\eta}(\ell)$, stays close to unity at every layer band of every non-excluded model (Table 2): the largest mean deviation in any cell is 5.5% (Gemma 2 2B at the final layer), and typical deviations are below 2%. Per-query spread is moderate at early and mid layers (SD typically 0.05–0.10) and grows at the final layer (SD up to 0.22 in Gemma 2 2B), but the mean remains near unity throughout: equal magnitudes are a property of the aggregate over queries—a population-level regularity—not of every individual query. The consequence is geometric. At the population mean, where $|\bar{\eta}(\ell) - 1| < \delta$ (Definition 1, Appendix B), the squared distance between correct and incorrect representations is dominated by the angular term $2\,\bar{r}_\ell^{\,2}(1 - \cos\theta_\ell)$; the magnitude term contributes only to the extent that individual queries depart from $\eta_i = 1$. The discriminative signal is therefore carried principally by the angle between displacement vectors rather than by their lengths, though at the level of an individual query a magnitude contribution remains.

4.2 Asymmetric Logit-Lens Commitment in the Incorrect Run

Figure 2 shows the logit-lens commitment ratio $\kappa(\ell)$ across normalized depth for the six measurable models; Table 3 reports per-model summary statistics with significance tests against the naive prior $\kappa = 0.5$. Solid colored lines show κ during the correct run, stratified by query category (deep constraint red, control blue, neutral grey); the dashed grey line shows the mean value of κ during the incorrect run. In LLaMA-2 13B, Mistral 7B and Llama 3.2 3B, correct-run commitment rises from 0.50 at the embedding layer to 0.76–0.79 at the final layer (Table 3), with deep-constraint queries (red) leading control (blue) throughout; across all six measurable models the final-layer correct-run commitment spans 0.72 (StableLM-2) to 0.93 (Gemma 2). The incorrect-run dashed line falls to $\kappa_{\min} = 0.08$ in LLaMA-2 13B and Mistral 7B and 0.13 in Llama 3.2 3B. In Gemma 2 2B, the incorrect-run $\kappa_{\min} = 0.08$ matches LLaMA-2 13B despite an order-of-magnitude smaller parameter count. StableLM-2 1.6B reaches $\kappa_{\min} = 0.32$; Llama 3.2 1B reaches $\kappa_{\min} = 0.25$. The observed $\kappa_{\min} = 0.08$ corresponds to the model assigning the incorrect token approximately 11.5 times the probability of the correct token, with κ in the incorrect run far below the naive prior of 0.5. The asymmetry is consistent across the four measurable architecture families (Llama, Mistral, Gemma, StableLM) and replicates within architecture families with scale.

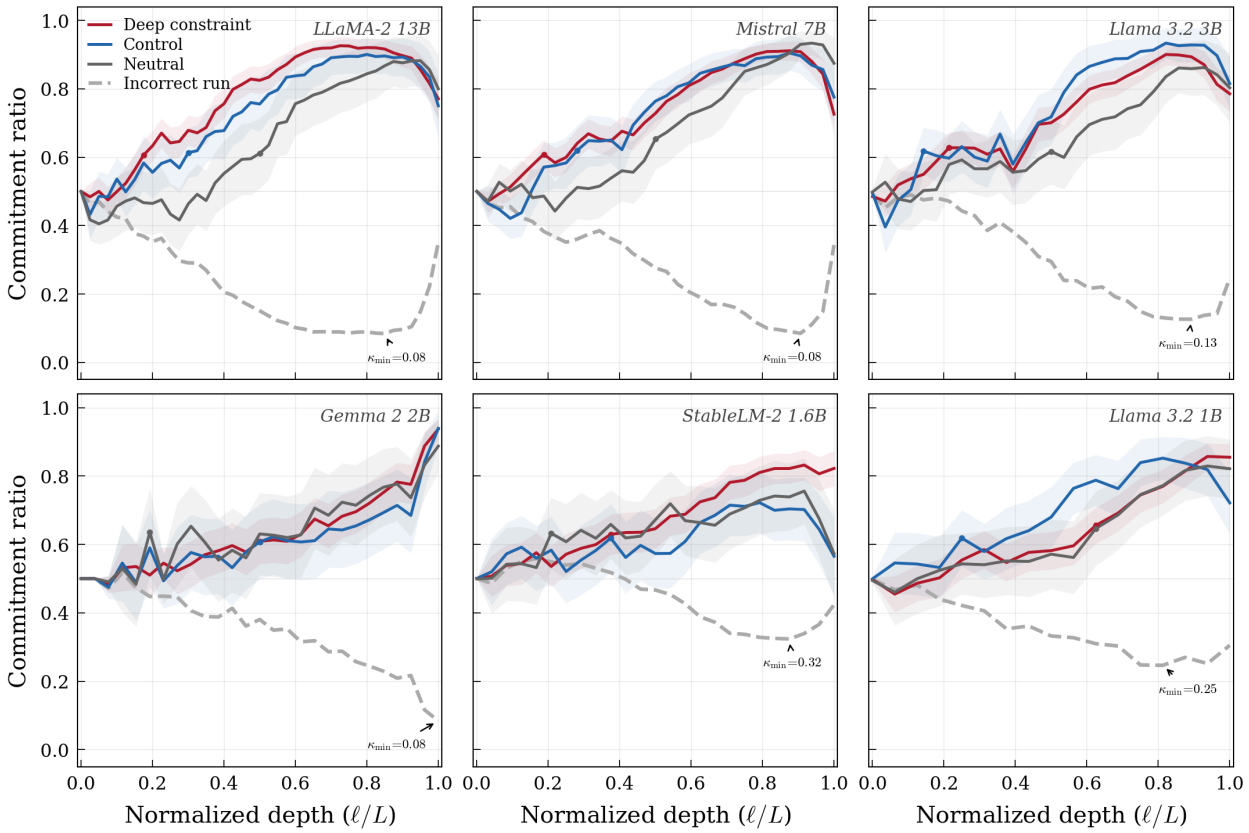


Figure 2: Logit-lens commitment dynamics across the six models with measurable factual processing (2×3 grid, ordered large to small). Solid lines: $\kappa(\ell)$ during correct-run processing, stratified by query category (DC red, CTRL blue, NEU grey). Dashed grey line: mean $\kappa(\ell)$ during incorrect-run processing. Horizontal line at $\kappa = 0.5$ marks the naive prior of equal probability assignment. Annotated $\kappa_{\min}$ values indicate the minimum commitment reached during incorrect-run processing. Qwen2 1.5B (not shown) holds $\kappa \equiv 0.50$ throughout under our extraction protocol; see Section 6.

4.3 Linear Probe Accuracy

We observe in Figure 3 that all six measurable models rise from chance at the embedding layer to a peak at intermediate depth ($\ell/L \in [0.21, 0.52]$) and then decline toward the final layer; within the Llama family peak accuracy scales with parameter count (0.74 at 1B, 0.78 at 3B, 0.85 at 13B), and Gemma 2 2B peaks at 0.78 matching the 3B Llama. The post-peak decline of $\Delta = 0.07$–0.12 is consistent across all six models. Full per-model probe accuracies are in Table 4 (Appendix B). The qualitative shape is consistent with prior layerwise probe work [Alain and Bengio, 2017, Belrose et al., 2023].

4.4 Attention Reallocates to Deep-Constraint Tokens in Correct Runs

The attention allocation ratio $\alpha(\ell)$ (M5) to manually annotated deep-constraint tokens, averaged across heads at the response-token position, is systematically higher in the correct run than in the incorrect run (Table 5, Appendix B). The per-model mean across queries and layers reaches $\bar{\alpha}^{+} = 0.60$–0.65 in correct runs versus $\bar{\alpha}^{-} = 0.42$–0.46 in incorrect runs. The ≈ 0.15–0.20 gap is significant at 15–39 of the available layers per model under FDR-corrected Mann-Whitney U

Table 3: Logit-lens commitment dynamics across the seven models. $\bar{\kappa}^{+}_{\text{final}}$: mean commitment at the final layer during correct-run processing. $\bar{\kappa}^{-}_{\text{min}}$: minimum mean commitment during incorrect-run processing. Suppression depth $\sigma = 0.5 - \bar{\kappa}^{-}_{\text{min}}$. One-sample t-test against $\kappa = 0.5$ at the minimum-κ layer; $n = 300$.

Model	$\bar{\kappa}^{+}_{\text{final}}$	$\bar{\kappa}^{-}_{\text{min}}$	ℓ^*/L	σ	t	p
LLaMA-2 13B	0.77	0.08	0.85	0.42	-36.5	$< 10^{-100}$
Mistral 7B	0.76	0.08	0.91	0.42	-39.1	$< 10^{-100}$
Gemma 2 2B	0.93	0.08	1.00	0.42	-30.1	$< 10^{-91}$
Llama 3.2 3B	0.79	0.13	0.93	0.37	-25.0	$< 10^{-74}$
StableLM-2 1.6B	0.72	0.32	0.88	0.18	-7.9	$< 10^{-13}$
Llama 3.2 1B	0.82	0.25	0.87	0.25	-12.8	$< 10^{-29}$
Qwen2 1.5B	0.50	0.50	–	0.00	-0.01	0.99

testing, with rank-biserial correlations $r_{\max} = 0.35$–0.43 across the six measurable models. Llama 3.2 1B reaches $r_{\max} = 0.42$ on attention while showing only moderate κ asymmetry. We report this correlation as an observation, not as evidence that attention reallocation causes the commitment asymmetry.

4.5 Cross-Domain Probe Transfer and Structural Asymmetry

We train linear probes on hidden states from one domain and test on each of the other two. The right panel of Figure 3 shows transfer AUROC across normalised depth. Within-domain AUROC is near-perfect for all six models from early mid-layers onward. Cross-domain transfer is consistently weaker and structurally asymmetric: financial–medical AUROC reaches 0.80–0.96 at the final layer, while transport-domain transfer to or from either other domain yields 0.52–0.74. The transport-domain asymmetry replicates across all six models, which makes a model-specific artefact unlikely; we read it as reflecting the shared regulatory-numeric structure of the financial and medical stems relative to the transport stems, and note that we do not have an independent measure of pretraining co-occurrence to confirm that reading.

4.6 Single-Layer Activation Patching

On LLaMA-2 13B and Mistral 7B—the two models where the patch is cleanly interpretable—single-layer activation patching produces no layer band where patching from the correct run recovers the correct token in the incorrect run. LLaMA-2 13B shows mean per-layer effect ≈ 0 across depth bands and category types. Mistral 7B shows a small, high-variance category-conditional effect on deep-constraint queries (mean 0.32 ± 0.96) but not on control or neutral, and the effect does not persist across the layer band; given the standard deviation it does not support a localized site. StableLM-2 1.6B instead recovers uniformly at every layer—including the first—with mean effect at or above full recovery (≈ 1.0, and 1.38 on control queries, i.e. beyond the recovery ceiling of 1). A valid single-layer patch cannot recover the answer from the input to the first layer, and normalized recovery cannot legitimately exceed 1; this profile is diagnostic of an invalid patch on the StableLM code path (Appendix D) rather than of localization. We therefore exclude StableLM-2 from the causal analysis pending a corrected re-run, and report a *scoped null*: within the two cleanly-patched architectures, single-layer patching does not localize the late-layer asymmetry. Per-layer effect tables are in Appendix D.

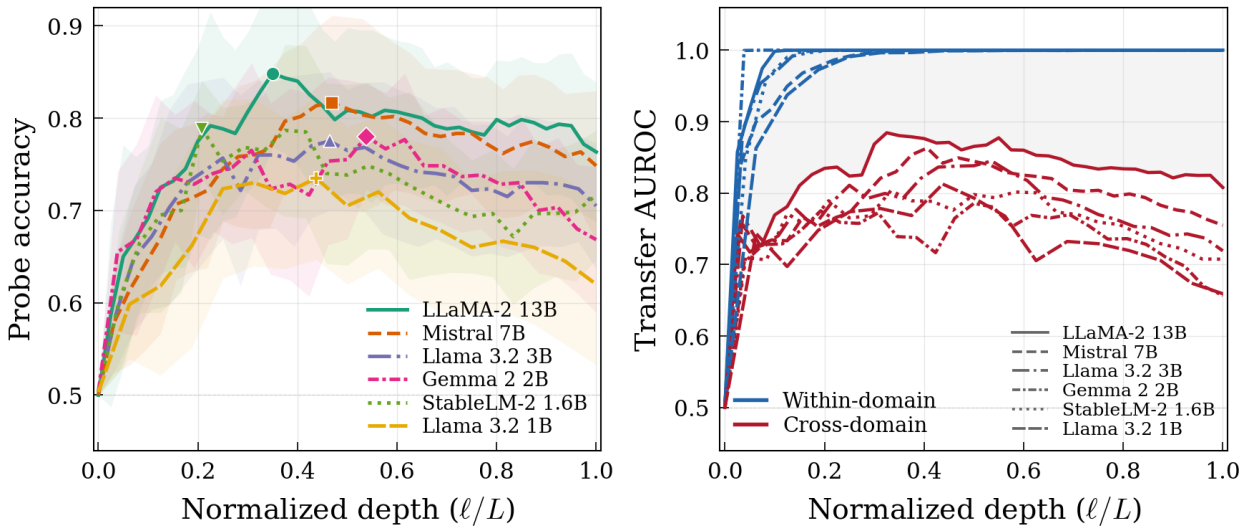


Figure 3: Probe accuracy and cross-domain transfer across the six measurable models. **Left:** Linear probe accuracy (5-fold CV, shaded ±1 SD) across normalised depth. All models peak at intermediate depth; within the Llama family the peak shifts deeper and rises with parameter count. **Right:** Cross-domain transfer AUROC. Blue lines: within-domain (near-perfect for all models from early mid-layers onward). Red lines: cross-domain. The financial–medical corridor (upper red lines) transfers consistently better than transport-domain pairs (lower red lines), a structural asymmetry replicated across all six models.

5 Discussion

5.1 Scope of the geometric claim

In the incorrect run, $\kappa_{\min} = 0.08$ means the model assigns the appended incorrect token roughly 11.5× the probability of the correct one at the response position. The matched-pair design controls for syntactic position and domain-meaningful framing of t^+ and t^-; under that design, a model treating both continuations as equally valid completions would produce $\kappa \approx 0.5$ in both runs. The observed asymmetry is therefore difficult to explain by neutral context-following alone. Two readings remain compatible with the incorrect-run trajectory: the model rejects the wrong continuation, or it conforms to the token it is forced to carry. Our claim rests on the divergence between the matched runs rather than on the incorrect run alone; separating rejection from conformity is exactly what a random-token baseline—a third forced run with a frequency-matched but semantically unrelated token—would do, and it is the priority next experiment (Section 6). The asymmetry calls for a causal explanation, but its mechanism remains open: the scoped null on single-layer patching (Section 4.6) bounds what causal account our data support on the two cleanly-patched architectures, and we cannot decide between a mechanism distributed across many small contributions and a localized mechanism that single-layer patching fails to detect. Discriminating these accounts further requires span-based patching across consecutive layers and direction-based interventions on identified residual-stream subspaces. We do not claim that our measurements characterize hallucination as it arises in free generation; the protocol is a controlled probe under forced continuations, and whether identical dynamics arise under unconstrained generation is the central empirical question we leave open.

5.2 Implications for measurement and detection

Four implications follow within the forced-completion regime and require verification under free-generation conditions before being extended. (i) Because divergence is isometric at tolerance $\delta \leq 0.06$ at the population mean (Definition 1, Table 2), methods that compare only displacement magnitudes will miss the primary signal; direction-based metrics are the relevant tool. (ii) Because κ asymmetry forms through mid-to-late depth rather than at the output, layerwise probes access information that output-layer detection cannot. (iii) Because cross-domain probe transfer is structurally asymmetric, per-domain calibration is prudent; the financial–medical corridor transfers far better than transport pairs, a regularity consistent with shared regulatory-numeric stem structure and with regional and topological accounts [Steenrod, 1951, Bronstein et al., 2021], but not adjudicated by our data. (iv) Because within the scales tested architecture rather than parameter count alone appears to set the suppression-depth ceiling—Gemma 2 2B reaches $\sigma = 0.42$ at 2B parameters, matching LLaMA-2 13B at 13B and exceeding the equivalent-scale Llama 3.2 3B ($\sigma = 0.37$)—within-family parameter sweeps are needed to separate the architectural contribution from scale. Candidate architectural choices include Gemma 2's alternating local/global attention and logit softcapping [Gemma Team, 2024]; we flag these as hypotheses, not conclusions, since a four-family sample cannot isolate a single mechanism.

6 Limitations

Scope of inference. The protocol observes hidden-state geometry under matched forced continuations, not under unconstrained autoregressive generation; we do not claim identity of dynamics across regimes. Free-generation experiments on the same query stems with model-produced hallucinations are the priority extension.

Causal protocols. Two controls are missing. (i) A random-token baseline (a third forced run with a tokenizer-frequency-matched but semantically unrelated continuation) would discriminate sensitivity to factual mismatch from general accommodation of any forced continuation—the rejection-vs-conformity ambiguity noted in Section 5. (ii) A query-only logit-lens trajectory would measure intermediate preference for the correct token before any continuation is forced. Together with span-based patching, direction-based interventions on residual-stream subspaces, resample ablations, and a corrected patching hook for the StableLM code path, they would discriminate distributed from localized-but-undetected mechanistic accounts.

Patching scope. Single-layer patching was run on three of the seven models and is cleanly interpretable on two (LLaMA-2 13B, Mistral 7B); the StableLM-2 result is excluded as an artefact of its block-container code path (Appendix D). The scoped null therefore rests on two architectures of the same broad family and should not be read as a claim about all seven models or all architectures.

Dataset and model scope. The 300 queries are human-authored under a fixed protocol by a single annotator; inter-annotator reliability on a 50-query partial re-categorization is planned. The forced-completion protocol assumes a single BPE response token; for Qwen2 1.5B several financial/medical responses fragment into sub-tokens, so the last-token extraction reads from a non-response position—this explains the flat κ and means Qwen2 1.5B's characterization as a scale floor is not safely supported until per-tokenizer extraction is corrected. Five architecture families across seven models do not separate scale from architectural effects cleanly; within-family parameter sweeps are needed, and encoder–decoder and retrieval-augmented architectures may differ.

7 Conclusions

We introduced forced-completion probing as a controlled instrument for measuring layerwise hidden-state geometry of factual queries, applied it to seven decoder-only transformers across five architecture families, and documented a near-isometric rotational divergence pattern together with an asymmetric late-layer commitment, alongside a scoped null result for single-layer causal localization on the two architectures where patching is cleanly interpretable. The geometric structure is observable and robust across six measurable models; the mechanism that produces it remains open, as does whether identical dynamics arise under unconstrained autoregressive generation, and whether the incorrect-run trajectory reflects rejection or conformity. We release the protocol, the 300-query annotated dataset, and the measurement battery to enable the random-token, query-only, corrected-patching and span/direction-based intervention experiments that would close these questions.

Reproducibility

The dataset, measurement code, and complete extraction protocol (including the per-tokenizer response-position localization and the architecture-specific patching-hook alignment discussed in Section 6 and Appendix D) are available under request to author. All experiments were run on a single NVIDIA A100 40 GB. Per-query, per-layer, per-model raw measurements are released as CSV files; aggregation scripts that produce the tables and figures in this paper are released as Jupyter notebooks.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. Stable LM 2 1.6B technical report. *arXiv preprint arXiv:2402.17834*, 2024.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (Volume 1: Long Papers)*, pages 8493–8502, 2022.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, 2023.

Alexander Yom Din, Noga Tamir, Idan Szpektor, and Yoav Goldberg. Jump to conclusions: Short-cutting transformers with linear transformations. *arXiv preprint arXiv:2303.09435*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65, 2019.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9484–9495, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 758–786, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Emmy Liu, Varun Gangal, Chelsea Zou, Michael Yu, Xiaoqi Huang, Alex Chang, Zhuofu Tao, Karan Singh, Sachin Kumar, and Steven Y. Feng. A unified definition of hallucination: It's the world model, stupid! *arXiv preprint arXiv:2512.21577*, 2026. Author list to be verified from arXiv before camera-ready.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2024.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Norman Steenrod. *The Topology of Fibre Bundles*. Princeton University Press, 1951.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

A Visual Summary

Figure 4 summarizes the joint observational pattern documented in Sections 4.1 and 4.2. We show this figure to highlight two simultaneous regularities. First, in Panel A, the dashed concentric arcs make explicit that the four displacement vectors at successive normalized depths terminate at approximately equal radii in the correct and incorrect runs—the radial differences between the two pathways are small relative to the angular separation at every depth. The discriminative signal is encoded in the angle $\theta(\ell)$, not in the magnitudes. Second, in Panel B, the dashed horizontal at $\kappa = 0.5$ is the naive prior an observer would expect if the residual stream merely accommodated

the appended token without preference. The $\kappa_{\min} = 0.08$ achieved in the incorrect run corresponds to the model placing approximately 11.5× more probability mass on the incorrect token than on the correct one—an asymmetry whose mechanistic explanation remains open under our scoped single-layer patching null (Section 4.6 and Appendix D).

B Supplementary Tables and Definitions

This appendix collects per-model numerical tables omitted from the main text for space, and the formal definition of *isometric divergence* referenced from Section 4.1.

Definition 1 (Isometric divergence)**.** *A pair of forced runs exhibits* isometric divergence *at layer ℓ* at tolerance δ *if $|\bar{\eta}(\ell) - 1| < \delta$. Under this condition, the squared distance between the mean correct and incorrect representations satisfies*

$$\|h_\ell^+ - h_\ell^-\|^2 \;=\; 2\,\bar{r}_\ell^{\,2}\,(1 - \cos\theta_\ell) + O(\delta), \tag{2}$$

where θ_ℓ is the angle between displacement vectors and $\bar{r}_\ell$ is the mean displacement magnitude across the two runs. At the population mean the discriminative signal resides predominantly in θ_ℓ rather than in the magnitudes; the magnitude contribution vanishes as $\delta \to 0$. The statement is a population-mean identity: for an individual query with $\eta_i \neq 1$, a magnitude contribution remains. In our data (Table 2), isometric divergence holds at tolerance $\delta \leq 0.06$ across all cells.

Table 4: Linear probe accuracy (5-fold cross-validation). All six measurable models peak at intermediate depth with a consistent post-peak decline.

Model	Peak accuracy	Peak ℓ/L	Final accuracy	Δ
LLaMA-2 13B	0.85 ± 0.08	0.35	0.76	0.09
Mistral 7B	0.82 ± 0.09	0.47	0.75	0.07
Llama 3.2 3B	0.78 ± 0.08	0.45	0.71	0.07
Gemma 2 2B	0.78 ± 0.06	0.52	0.67	0.11
StableLM-2 1.6B	0.79 ± 0.05	0.21	0.72	0.07
Llama 3.2 1B	0.74 ± 0.09	0.41	0.62	0.12
Qwen2 1.5B	0.50	–	0.50	0.00

Table 5: Attention allocation to deep-constraint tokens across the seven models. Correct runs systematically allocate more attention to factual anchor tokens. FDR-corrected Mann-Whitney U test across layers; r is rank-biserial correlation.

Model	$\bar{\alpha}^+$	$\bar{\alpha}^-$	Sig. layers	$r_{\max}$	Best p_{FDR}
LLaMA-2 13B	0.62	0.46	39/41	0.35	$< 10^{-5}$
Mistral 7B	0.60	0.42	31/33	0.35	$< 10^{-5}$
Llama 3.2 3B	0.64	0.45	28/29	0.41	$< 10^{-6}$
Gemma 2 2B	0.63	0.42	21/27	0.43	$< 10^{-6}$
StableLM-2 1.6B	0.61	0.42	18/25	0.40	$< 10^{-6}$
Llama 3.2 1B	0.65	0.42	15/17	0.42	$< 10^{-5}$
Qwen2 1.5B	0.48	0.34	0/29	0.07	0.20

C Dataset Construction

This appendix expands the summary in Section 3.2. The dataset is 300 human-authored queries, stratified across three professional domains ($n = 100$ each) and three category types under a fixed authoring protocol.

Deep constraint (DC, $n = 182$). The correct answer requires integrating a domain-specific factual rule with the query context. The query stem contains both a *deep-constraint cue* (factual content that determines the correct answer under domain knowledge) and a *surface cue* (a syntactic or numerical pattern that, taken alone, would favor the incorrect answer). The per-domain DC categories are:

- **Financial compliance:** *RA* (risk-aware suitability: retirement-age portfolio rules, conservative-investor risk profiles) and *RC* (regulatory conflict: leverage prohibitions for conservative investors, suitability under FINRA/MiFID-style frameworks).
- **Medical protocols:** *DI* (drug interactions: e.g., warfarin-aspirin, SSRI-NSAID) and *CI* (contraindications: e.g., methotrexate-NSAID, beta-blocker–asthma).
- **Transport regulation:** *FD* (functional dependency: the object must be at the destination for the task, e.g., taking a car to a car wash) and *PT* (physical transport: object too heavy or bulky to carry on foot regardless of distance).

Control (CTRL, $n = 66$). The correct answer is determinate and the surface and deep cues agree. The protocol matches the deep constraint categories in length and surface complexity but removes the surface-vs-deep tension. Per-domain categories are *FC* (financial), *MC* (medical), and *TD* (transport—distance is the legitimate basis for the decision).

Neutral (NEU, $n = 52$). The correct/incorrect distinction does not depend on domain-specific factual reasoning, serving as a within-protocol null control following Hewitt and Liang [2019]. Per-domain categories are *FN*, *MN*, *ND*.

Authoring constraints. Each query is human-authored under three constraints: (i) exactly one single-token correct continuation t^+ and exactly one single-token incorrect continuation t^-, both domain-meaningful and not arbitrary distractors; (ii) for the studied tokenizers, both t^+ and t^- map to single BPE tokens (per-model tokenizer constraints in Section 6); (iii) deep-keyword and surface-keyword annotations are character-offset-aligned to the query string for attention analysis.

Example query (RA, financial domain). `Query:` "Client is 68 and retired with $500K savings. High-yield bonds at 8% or treasury bonds at 4%?" `Correct:` "Treasury bonds. High-yield bonds carry significant default risk inappropriate for a retiree's portfolio." `Incorrect:` "High-yield bonds. Double the return means faster growth of retirement savings." `Deep keywords:` retired, default, risk, inappropriate, retiree. `Surface keywords:` 8%, return, faster, growth.

Example query (DI, medical domain). `Query:` "Patient has a mild headache and is on warfarin. Should they take aspirin or acetaminophen?" `Correct:` "Acetaminophen. Aspirin interacts with warfarin and significantly increases bleeding risk." `Incorrect:` "Aspirin. It is the standard and most effective headache treatment." `Deep keywords:` warfarin, interact, bleeding, aspirin. `Surface keywords:` mild, headache, standard.

Example query (FD, transport domain). `Query:` "I want to wash my car and the car wash is only 100m away. Should I drive or walk?" `Correct:` "Drive. You need the car at the car wash." `Incorrect:` "Walk. It is only 100m away, no reason to drive such a short distance." `Deep keywords:` wash, car. `Surface keywords:` 100m, away. The complete dataset, including per-query annotations of deep-keyword and surface-keyword character spans, is released alongside the paper.

D Activation Patching: Full Results and Protocol Note

This appendix expands Section 4.6. Table 6 reports the mean per-layer patching effect across queries, normalized-depth bands, and category types. The StableLM-2 row is retained for transparency but flagged as an artefact and excluded from interpretation (see below).

Table 6: Single-layer activation patching results. Effect $e(\ell)$ is the normalized recovery of the correct-token logit when the residual stream at layer ℓ is patched from the correct into the incorrect run; $e = 1$ indicates full recovery, $e = 0$ no causal contribution, $e < 0$ that the layer's correct-run state interferes with the incorrect run. Values are mean ± standard deviation across queries within each bin. StableLM-2 1.6B is flagged as an artefact of its block-container code path (see protocol note) and excluded from interpretation.

Model	Mean effect by depth band			Mean effect by category		
	Early	Mid	Late	DC	CTRL	NEU
LLaMA-2 13B	0.01 ± 0.98	-0.01 ± 0.97	-0.05 ± 0.95	0.02 ± 0.94	-0.03 ± 1.07	-0.12 ± 0.91
Mistral 7B	0.23 ± 0.98	0.21 ± 0.97	0.16 ± 0.96	0.32 ± 0.96	0.07 ± 1.00	-0.06 ± 0.90
StableLM-2 1.6B *(art.)*	1.04 ± 0.85	1.13 ± 0.80	1.01 ± 0.89	0.95 ± 0.82	1.38 ± 0.74	1.14 ± 0.91

Three observations. First, in LLaMA-2 13B, mean per-layer effect is approximately zero across all depth bands and all category types. No layer or layer band emerges as a localized causal site for correct-token recovery in the incorrect run. Second, in Mistral 7B there is a small category-conditional effect: mean effect on deep-constraint queries is 0.32 compared to −0.06 on neutral. The effect is in the expected direction but small in magnitude, high in variance (±0.96), and inconsistent across the layer band, so it does not support a localized site. Third, in StableLM-2 1.6B mean patching effect is at or above full recovery (≈ 1.0, and 1.38 on control) uniformly across layers—including the earliest—and across category types. Because a single-layer patch at the first layer cannot supply the downstream computation that produces the answer, and because normalized recovery cannot legitimately exceed 1, this profile is not interpretable as localization; it is diagnostic of an invalid patch on the StableLM code path. We exclude StableLM-2 from the causal conclusion.

Interpretive bound. We are deliberately conservative about the consequences of these results. The scoped null (on LLaMA-2 13B and Mistral 7B) does not establish that the late-layer asymmetry is distributed across many components rather than localized in a way our protocol fails to detect. Span-based patching (patching across multiple consecutive layers), direction-based interventions on identified residual-stream subspaces, and resample ablations would all be needed to discriminate between distributed-processing and localized-but-undetected accounts.

Implementation note on hook recursion. We document the following implementation detail because earlier protocol versions failed silently. Under `transformers` ≥ 4.40, setting `output_hidden_states=True`

installs internal forward hooks on every layer to collect hidden states. If our patching protocol additionally installs `forward_pre_hook`s on the same layers, the result is mutual recursion between our hook and the internal collection hook, manifesting either as infinite-loop failure (caught by the Python recursion limit) or, more dangerously, as silent NaN output. The fix used in our protocol is to set `output_hidden_states=False` at the model call site and pass hidden-state collection requests transiently per forward pass. We install our pre-hook on one layer at a time with a hook handle that is always removed in a `finally` block, regardless of forward-pass exceptions. The released code documents this in inline comments.

Architecture-specific source alignment. A second implementation detail explains the StableLM-2 artefact. Our patch replaces the residual-stream input to layer ℓ in the incorrect run with the source state collected at the corresponding index from the correct run. The index alignment between the collected hidden-state sequence and the module receiving the pre-hook was validated for models exposing the `model.model.layers` block container (LLaMA-2, Mistral, and the Llama 3.2 models). StableLM-2 exposes a different container (`transformer.h`); under that path the source index does not align to the same residual position, so the injected state carries near-final information at every layer, which produces the observed uniform, above-ceiling recovery. The corrected implementation we release captures the source and applies the patch through the identical `forward_pre_hook` mechanism, so alignment holds by construction across architectures; re-running StableLM-2 (and adding Gemma 2 and the smaller models) under the corrected hook is left to the intervention experiments described in Section 6. Until then we restrict the causal claim to the validated code path.

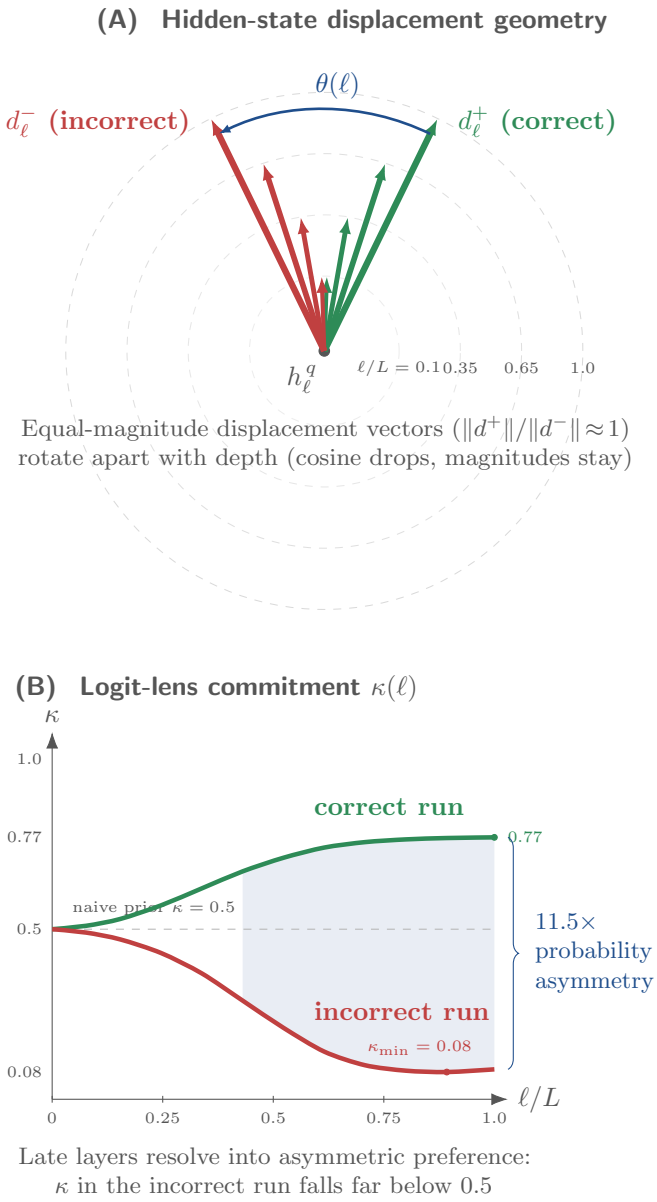


Figure 4: Visual summary of the central observational pattern. **(A)** In displacement space, correct- and incorrect-run vectors from the query-only representation h_ℓ^q rotate apart with depth while maintaining approximately equal magnitude ($\bar{\eta}(\ell) \approx 1$; see Table 2). The angular separation $\theta(\ell)$ grows through mid-depth before partial reconvergence at the final layer. **(B)** Late layers resolve the asymmetric outcome: the logit-lens commitment ratio $\kappa(\ell)$ rises above the naive prior of 0.5 in the correct run and falls far below it in the incorrect run, reaching $\kappa_{\min} = 0.08$ in the largest models (Table 3). The asymmetry corresponds to the model assigning the incorrect token approximately 11.5× the probability of the correct token at the minimum-κ layer.