# BEFORE THE FIRST TOKEN: LOCALIZING AND CONTROLLING PRE-GENERATION RESPONSE-LENGTH PLANS IN A LANGUAGE MODEL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Autoregressive language models generate one token at a time, yet their responses often exhibit global properties that appear to be determined before generation begins. We study one measurable instance of this phenomenon: response length. Using Llama-3.1-8B-Instruct, we first show that substantial target-dependent length control persists when an explicit length instruction is visible during prefill but masked throughout decoding. Clean-recipient activation patching localizes a dominant transferable pathway to fixed post-user template positions and to a distributed mid-layer residual trajectory. A counterfactual prefix experiment further shows that this effect is not explained primarily by the identities of the first 1–32 generated tokens. We then introduce a cross-replicated estimator that extracts length-aligned activation trajectories using within-prompt variation in actual response length while retaining only structure that replicates across instruction paraphrases. The resulting layer-specific low-rank spaces replicate on independent prompts. Finally, on 120 unseen natural prompts containing no length instruction, prefill-only interventions within the frozen spaces bidirectionally and dose-dependently change response length, outperform matched random-subspace controls and remaining effective when the first 16 response tokens are fixed. These results support a pre-generation causal representation of response length in this model, while leaving precise target control, semantic-quality preservation, and cross-model generality unresolved.

## 1 INTRODUCTION

An autoregressive language model (LM) exposes only a local next-token distribution, but a useful response must satisfy constraints that extend far beyond its next token. It may need to remain concise, develop an argument over several paragraphs, or terminate near a requested budget. Length is a particularly useful probe of such global organization: it is objective, varies naturally across tasks, and cannot generally be implemented by changing a single lexical choice. Instruction-following benchmarks confirm that length constraints are behaviorally meaningful and often difficult (Zhou et al., 2023; Zhang et al., 2025). What remains unclear is whether an LM forms an internal length plan before generating the response, where such a plan is represented, and whether that representation itself has a causal effect on generation.

There are at least three competing explanations for apparent length planning. First, the model may reread an explicit length instruction throughout decoding, without committing to a pre-generation state. Second, the instruction may merely change the opening tokens, after which ordinary autoregressive dynamics produce a different length. Third, hidden states may correlate with eventual length because of prompt semantics or instruction wording while carrying no causal control. Linear probing alone cannot distinguish these explanations. We therefore organize the study as a sequence of interventions that progressively narrows the claim.

We first establish and localize a *pre-generation causal state*, without assuming that it is linear or low-dimensional. Exact attention masking removes decode-time access to the length instruction; clean-recipient residual patching then asks which prefill positions and layers can transfer control into



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.

a context that has never seen that instruction. We additionally exchange early generated token IDs between clean and length-conditioned contexts to separate prefill-state effects from response-opening effects. These experiments identify the post-user template and middle residual depths as a dominant, but not necessarily exclusive, transferable pathway.

We next ask whether the identified states contain a shared geometry across prompts. Simple PCA or mean-difference directions can be dominated by prompt semantics and by instruction surface form. We instead estimate response-length trajectories only from variation within a fixed prompt, instruction family, and paraphrase, and construct a cross-replicated operator from independently worded condition sets. This yields a dominant shared axis with low-rank prompt-specific refinements. The resulting spaces are frozen before evaluation on independent prompts and before causal intervention.

Finally, we intervene on unseen natural prompts that contain no length request and apply no masking during decoding. Signed, prefill-only changes within the extracted spaces shorten and lengthen responses in a dose-dependent manner. Random, orthogonal-complement, full-state, layer-window, and fixed-response-prefix controls test specificity and rule out several disruption-based explanations.

Our contributions are:

- We provide causal evidence that target-dependent response-length information is present before the first generated token, and localize a dominant transferable pathway to post-user template states across middle layers.
- We introduce a cross-replicated trajectory estimator that separates actual-length-aligned variation from prompt offsets and instruction-paraphrase-specific variation.
- We show that the frozen representation causally modulates response length on 120 unseen natural prompts, with bidirectional, dose-responsive effects that exceed matched controls and persist under a fixed response opening.

We deliberately limit the headline claim to Llama-3.1-8B-Instruct. Cross-model replication and quality-preserving precise target control are important next steps rather than established results.

## 2 Related Work

**Instruction following and length control.** Instruction tuning enables language models to follow diverse natural-language constraints (Ouyang et al., 2022), and verifiable evaluations include minimum- and maximum-length requirements (Zhou et al., 2023). Recent work documents persistent difficulty at large requested lengths (Zhang et al., 2025) and develops architectural or fine-tuning mechanisms for precise length control (Butcher et al., 2024). In contrast, we do not train a controller or evaluate only instruction compliance. We use length as an observable variable for studying a pre-generation internal mechanism.

**Planning before generation.** Transformers can place information in current hidden states that is useful at later positions, although whether this reflects dedicated pre-computation or features useful both now and later remains debated (Wu et al., 2024). Our experiments address a narrower behavioral quantity. By masking the instruction after prefill and fixing generated prefixes, we test whether a response-level property is causally supported before free decoding begins.

**Causal localization and representation intervention.** Activation patching and causal tracing intervene on internal states to localize model behavior (Meng et al., 2022; Zhang & Nanda, 2024). Representation engineering and activation addition instead seek population-level directions that can monitor or steer behavior (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2024). These methods motivate our use of residual-stream interventions, but also raise two concerns: a direction can mix target semantics with nuisance variation, and a subspace intervention can activate a pathway that is not used naturally (Makelov et al., 2024). We address the first with within-cell estimation and cross-replication, and the second with natural-prompt interventions, matched subspace controls, and full-state comparators.

# 3 EXPERIMENTAL SETUP

**Model and prompts.** We study Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with its native chat template. Prompts are sampled from UltraChat (Ding et al., 2023) and filtered to exclude existing length, word-count, sentence-count, concision, or detail requirements. All reported generations use greedy decoding with a limit of 1,800 new model tokens. Length requests are expressed in words, matching ordinary user instructions; all outcomes count generated model tokens after removing the terminal EOS token. We never compute an MAE between requested words and generated tokens because those quantities have different units.

**Positions and depths.** We call the fixed tokens between the user message and the first response token the *post-user template*. Its five positions are denoted C1–C5. Unless stated otherwise, extraction and subspace intervention use C5. Residual depth $h_\ell$ is the input to zero-indexed Transformer block $\ell$ ("resid_pre"); thus $h_{11}$ is the output of block 10. A patch at $h_\ell$ is processed by blocks $\ell, \ldots, 31$.

**Data partitions.** The localization experiments use 150 prompts for masking and prefix mediation, with a 51-prompt subset for clean-recipient layer discovery. Geometry extraction retains 48 training and 18 validation prompts after complete-cell quality control. A separate 100-prompt cohort provides 63 eligible prompts for exact-method geometry confirmation. The final causal experiment uses 20 development prompts for dose calibration and 120 confirmatory prompts disjoint from all prior prompt registries. Conditions for one prompt never cross partitions.

**Uncertainty and quality auditing.** The prompt is the statistical unit. We report prompt-bootstrap 95% confidence intervals (CIs). Confirmatory causal tests use two-sided prompt-level sign-flip tests with Holm correction over five predeclared hypotheses. Empty, truncated, refusal-like, and abnormally repetitive generations are audited separately. Because these automatic checks do not establish correctness or relevance, we do not claim preservation of semantic quality.

# 4 LOCALIZING A PRE-GENERATION CAUSAL STATE

## 4.1 LENGTH CONTROL PERSISTS AFTER DECODE-TIME MASKING

For a user instruction $x_i$ and requested word target $T$, the model first performs a normal prefill on the complete input. During generation, we set attention to the exact length-instruction span to zero at every decode step, leaving the user instruction, post-user template, and generated tokens visible. The first generated token is sampled from the original prefill logits; subsequent tokens cannot re-read the length request. We compare this *masked* condition with a visible-instruction condition and a no-length-instruction baseline.

Across 150 prompts and targets $T \in \{100, 300, 500, 800, 1000\}$ words, the suffix-masked response curve remains strongly ordered by target (mean prompt Spearman $\rho = 0.887$). Our preregistered 100-to-800 contrast spans 384.4 generated tokens under masking, retaining 47.9% [44.8, 51.3] of the visible-condition span. The curve saturates beyond 800 requested words, so the 1000-word point is descriptive. Moving the request before the user content raises masked retention to 56.1%; therefore we do not claim that all relevant state is confined to the post-user template. Rather, the result shows that substantial target sensitivity survives solely through states formed during prefill.

## 4.2 CLEAN-RECIPIENT PATHWAY LOCALIZATION

Masking demonstrates persistence but does not identify the responsible states. We therefore use a clean recipient containing only the user instruction and post-user template, and an aligned donor that also contains "Please answer in $T$ words." At a selected residual depth, we copy donor states at either all user-instruction positions, C1–C5, or both into the clean recipient. Every unpatched recipient position has never processed a length instruction. Downstream blocks recompute K/V in the recipient context, avoiding manual positional correction.

At the strongest point of an initial six-depth screen ($h_{12}$), post-user patching transfers a bidirectional 157.8-token 100-to-800 span [130.1, 186.8], or 36.7% of the full masked-instruction span. Patching

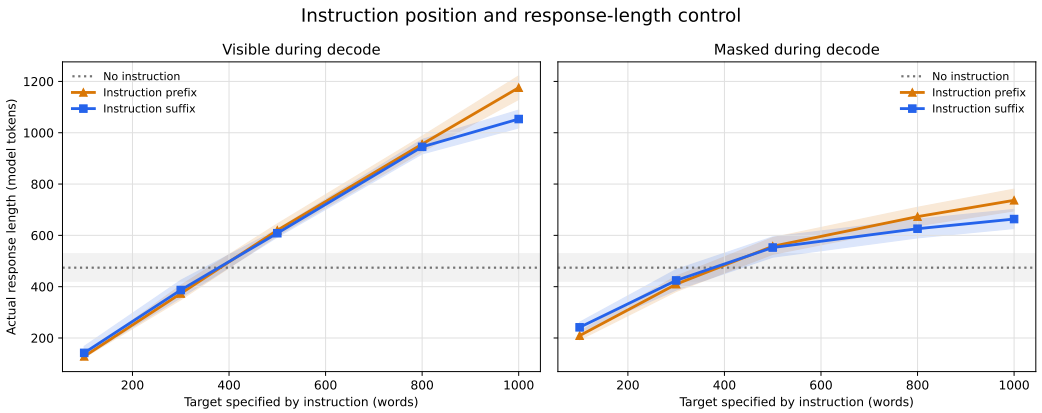



Figure 1: Response length as a function of a requested word target when the length instruction is visible throughout generation or masked after prefill. Targets are measured in words and responses in model tokens. Masked curves remain target-sensitive, showing that prefill-induced states support length control without decode-time access to the request.

all user-instruction positions transfers only 12.4 tokens $[-5.8, 31.1]$. Patching both regions gives 160.2 tokens, statistically indistinguishable from post-user patching alone. The larger number and residual norm of patched user states cannot explain this difference. We interpret C1–C5 as a dominant transferable terminal pathway, not as the unique location at which length information originates.

### 4.3 The pathway is distributed across middle depths

A complete single-depth scan localizes a peak plateau to $h_{10}$–$h_{11}$: at $h_{11}$, C1–C5 patching recovers 50.1% [42.1, 58.7] of the full masked span. Single-depth localization does not imply single-layer implementation, because subsequent blocks continue to transform the patched state. We therefore repeatedly clamp the donor C1–C5 states across preregistered depth windows. As shown in Figure 2a, $h_8$–$h_{15}$ recovers 91.6% [85.2, 98.6] of the masked effect, whereas equal-width early ($h_0$–$h_7$) and late ($h_{20}$–$h_{27}$) controls remain near zero. This discovery experiment motivates a layer-specific mid-layer trajectory rather than one state copied identically across depths.

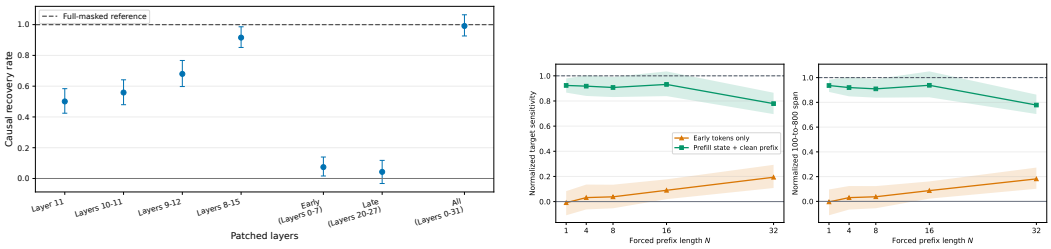
### 4.4 Early response tokens do not explain the prefill effect

The length-conditioned prefill could act only by selecting a different response opening. To test this, we define clean and masked-instruction contexts $C_0, C_T$, and their naturally generated first $N$ token IDs $Z_0, Z_T$. We compare early tokens only ($C_0 + Z_T$) against prefill state plus a clean prefix ($C_T + Z_0$). Token IDs are teacher-forced and all hidden states and K/V entries are recomputed in the recipient context; no donor cache is copied.

For $N \in \{1, 4, 8\}$, $C_0 + Z_T$ recovers at most 3.8% of masked target sensitivity, while $C_T + Z_0$ retains 90.7–92.4%. At $N = 32$, where every prompt has target-dependent 100/800 prefixes, the respective fractions are 19.3% and 79.1%. Thus, early token identity increasingly participates as decoding proceeds but is not the primary explanation for the masked length effect.

## 5 Extracting a Cross-Replicated Length Trajectory

The previous section establishes a causal pre-generation state but not a linear subspace. We now use controlled variation to extract a candidate geometry. For each prompt, 24 length-control conditions cross four instruction families (numeric, qualitative, structural, and bounded), two independently worded paraphrase sets $p \in \{A, B\}$, and short/medium/long induced responses. The no-length baseline is excluded from extraction.




(a) Residual-depth localization

(b) Response-prefix mediation

Figure 2: Controls localizing the pre-generation pathway. (a) Repeated clean-recipient patching across middle depths recovers most of the masked-instruction effect; equal-width early and late windows do not. (b) Target-derived response prefixes transfer little control to a clean context, whereas a target-conditioned prefill remains effective when clean response tokens are forced. Bands and error bars are prompt-bootstrap 95% CIs.

### 5.1 Within-cell actual-length trajectories

Let $\boldsymbol{h}_{ifpt}^{(\ell)} \in \mathbb{R}^d$ be C5 at depth $\ell$ for prompt $i$, instruction family $f$, paraphrase $p$, and tier $t$. Our label is the realized length $y_{ifpt} = \log(1 + L_{ifpt})$, not the requested target. We construct nested polynomial bases $\boldsymbol{\phi}_q(y)$, $q \in \{1, 2, 3\}$, whose normalization and orthogonalization are fit on training data. Within every $(i, f, p)$ cell, we subtract the tier means of both $\boldsymbol{h}$ and $\boldsymbol{\phi}$. This removes arbitrary prompt- and wording-specific offsets while preserving changes associated with realized length. A family is retained only when all short/medium/long conditions are clean for both A and B.

For each prompt and paraphrase, we concatenate complete families and estimate a regularized coefficient matrix

$$\boldsymbol{C}_i^p = \widetilde{\boldsymbol{H}}_i^{p\top} \widetilde{\boldsymbol{\Phi}}_i^p \left( \widetilde{\boldsymbol{\Phi}}_i^{p\top} \widetilde{\boldsymbol{\Phi}}_i^p + \lambda \boldsymbol{I} \right)^{-1} \in \mathbb{R}^{d \times q}. \tag{1}$$

Its first column is a local length direction; additional columns capture reproducible curvature. This is not a raw-state probe: only within-cell differences enter Equation 1.

### 5.2 Cross-replicated operator

We retain only coefficient structure that covaries across independently worded conditions:

$$\boldsymbol{K}_\ell = \frac{1}{2N} \sum_{i=1}^{N} \left( \boldsymbol{C}_i^A \boldsymbol{C}_i^{B\top} + \boldsymbol{C}_i^B \boldsymbol{C}_i^{A\top} \right). \tag{2}$$

$\boldsymbol{K}_\ell$ is symmetric but not necessarily positive semidefinite. We define the candidate space $\boldsymbol{U}_\ell$ from its leading positive eigenvectors and retain negative eigenvalues for audit. Unlike ordinary coefficient SVD, Equation 2 rewards geometry that replicates across A/B; unreplicated paraphrase-specific components cancel in expectation.

We select $(q, \lambda)$ on validation prompts using the mean of symmetric cross-paraphrase and leave-one-family-out normalized reconstruction error. We then select the smallest operational rank within one prompt-level standard error of the best validation loss. Five hundred within-cell label permutations provide a null spectrum. This rank is a regularized working dimension, not a claim about the model's true mechanistic dimensionality.

At $h_{11}$, one-SE selection gives $q = 2$, $\lambda = 1$, and rank 2. Rank 1 misses the one-SE threshold by only $7.3 \times 10^{-5}$ loss, so we describe the geometry as a dominant global axis with a small low-rank refinement and retain rank 1 as an ablation. On the earlier held-out development split, rank-2 cross-replication attains cross-paraphrase $R^2 = 0.578$ and cross-family $R^2 = 0.474$. Ordinary coefficient SVD obtains 0.541/0.529, while a direct prompt-slope SVD obtains 0.495/0.252. Thus cross-replication improves paraphrase reconstruction and substantially improves over the simple slope baseline across families, but it does not uniformly dominate coefficient SVD on every metric.

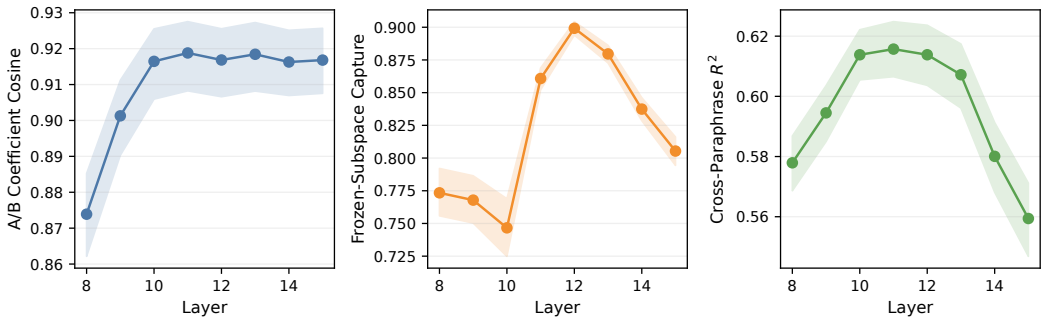



Figure 3: Independent confirmation of the frozen cross-replicated estimator across middle layers. Left: agreement between coefficient trajectories estimated from paraphrase sets A and B. Middle: fraction of confirmation coefficient energy captured by the frozen training space. Right: cross-paraphrase reconstruction in the frozen space. Lines are prompt means and bands are 95% bootstrap CIs. Rank and regularization are not reselected on these prompts.

### 5.3 Independent Geometry Confirmation

We freeze every layer's $(q, \lambda, r)$ before analyzing a separate cohort. Of 100 collected prompts, 63 satisfy the same complete-cell quality rule. At $h_{11}$, A/B coefficient reliability is 0.919 [0.908, 0.928], the frozen rank-2 space captures 0.861 [0.852, 0.869] of coefficient energy, and projected A/B agreement is 0.986 [0.981, 0.990]. An independently re-estimated confirmation space aligns with training at rank-1 cosine 0.990 and rank-2 projection similarity 0.955 (largest principal angle $15.9°$). The observed capture exceeds both within-cell label permutations ($p = 0.005$) and rank-matched ambient random spaces ($p = 0.002$).

Figure 3 shows that paraphrase reliability remains high across $h_8$–$h_{15}$, whereas capture and reconstruction vary by depth. Each depth has its own $\boldsymbol{U}_\ell$: the adjacent rank-1 axes have mean cosine 0.692, while the $h_8/h_{15}$ cosine is 0.168. We therefore interpret the signal as progressively re-encoded across layers, not as one invariant vector copied through the network.

## 6 Causal Validation on Unseen Natural Prompts

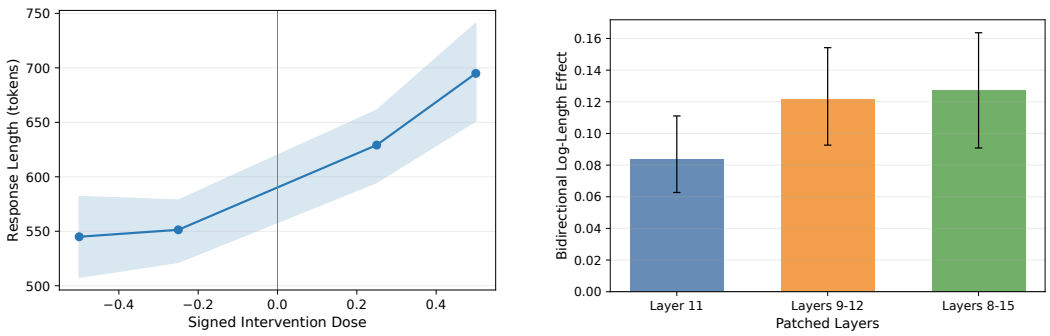
Geometry alone does not establish that the extracted space participates in natural generation. We therefore freeze all extraction choices and intervene on 120 unseen prompts containing no length instruction. No tokens are masked during decoding. For each $\ell \in \{8, \ldots, 15\}$, let $\boldsymbol{P}_\ell = \boldsymbol{U}_\ell \boldsymbol{U}_\ell^\top$. We estimate a prompt-balanced training displacement $\boldsymbol{g}_\ell = \boldsymbol{P}_\ell(\bar{\boldsymbol{h}}_{\text{long}} - \bar{\boldsymbol{h}}_{\text{short}})$ and modify only C5 during prefill:

$$\boldsymbol{h}'_{\ell,\text{C5}} = \boldsymbol{h}_{\ell,\text{C5}} + \alpha \boldsymbol{g}_\ell. \tag{3}$$

The orthogonal complement remains unchanged, and hooks are removed before the first decode step. We apply Equation 3 separately at each depth in a window, rather than transporting one vector across residual spaces. Twenty development prompts select a primary absolute dose of 0.5 under a preregistered safety rule; the choice and test manifest are then frozen.

### 6.1 Dose Response and Bidirectional Control

The four confirmatory doses $\{-0.5, -0.25, 0.25, 0.5\}$ produce aggregate mean lengths 545.0, 551.3, 629.1, and 694.9 tokens, respectively (Figure 4a). Relative to the natural baseline mean of 597.4 tokens, $\alpha = -0.5$ changes length by $-52.4$ tokens and moves 74.2% of prompts in the intended direction; $\alpha = 0.5$ changes it by $+97.6$ tokens and moves 80.0% in the intended direction. The prompt-level log-length dose slope is 0.251 [0.184, 0.318]. Separate short-directed and long-directed effects are both positive after Holm correction (Table 1), ruling out an intervention that merely shortens or lengthens all responses through generic disruption.




(a) Signed intervention dose

(b) Layer-window ablation

Figure 4: Causal intervention on 120 unseen natural prompts with no length instruction. (a) Mean response length changes monotonically with signed dose in the frozen layer-specific trajectory. (b) Repeated intervention across middle layers is stronger than intervention at $h_{11}$ alone; $h_9$–$h_{12}$ and $h_8$–$h_{15}$ are not detectably different at the frozen dose. Bands and error bars are prompt-bootstrap 95% CIs.

Table 1: Preregistered causal tests on 120 unseen natural prompts. Effects are computed from paired changes in $\log(1 + \text{response tokens})$. Reported $p$-values are Holm adjusted.

| Test | Mean effect | 95% CI | $p_{\text{Holm}}$ |
|---|---|---|---|
| Dose-response slope | 0.2511 | [0.1841, 0.3183] | $1.0 \times 10^{-4}$ |
| Short-directed effect | 0.1032 | [0.0388, 0.1684] | $4.2 \times 10^{-4}$ |
| Long-directed effect | 0.1516 | [0.1058, 0.2062] | $1.0 \times 10^{-4}$ |
| True minus random spaces | 0.1240 | [0.0790, 0.1753] | $1.0 \times 10^{-4}$ |
| Effect with first 16 tokens fixed | 0.1167 | [0.0918, 0.1441] | $1.0 \times 10^{-4}$ |

### 6.2 Specificity controls

The true $h_8$–$h_{15}$ trajectory has a mean bidirectional log effect of 0.1274 [0.0912, 0.1644]. Three independently seeded, rank- and displacement-norm-matched random subspaces have effects $-0.0009$, $-0.0289$, and 0.0400 (mean 0.0034). A rank-matched direction in the frozen space's orthogonal complement has a smaller but nonzero effect of 0.0456 [0.0189, 0.0764]; the true trajectory exceeds it by 0.0818 [0.0419, 0.1186]. Thus the extracted space contains the dominant targeted effect but does not exhaust all length-relevant variation. Full-state trajectory clamping gives 0.1510 [0.1174, 0.1859], so the frozen subspace recovers approximately 84% of that descriptive upper comparator on the log-length scale.

### 6.3 Layer distribution and fixed-prefix replication

Intervening only at $h_{11}$ yields a bidirectional effect of 0.0839 [0.0627, 0.1103]. The effect rises to 0.1219 [0.0930, 0.1549] for $h_9$–$h_{12}$ and 0.1274 [0.0912, 0.1644] for $h_8$–$h_{15}$. The broad window exceeds $h_{11}$ by 0.0435 [0.0104, 0.0751], but does not detectably exceed $h_9$–$h_{12}$ (difference 0.0055 $[-0.0173, 0.0260]$). This independently supports a distributed middle-layer mechanism without establishing that every layer in the broad window is necessary.

Finally, we teacher-force each prompt's first 16 natural baseline tokens under both intervention signs. The bidirectional effect remains 0.1167 [0.0919, 0.1452], and its difference from free decoding is only 0.0107 $[-0.0176, 0.0357]$. The extracted trajectory therefore continues to affect total length after the response opening is held fixed.

## 7 DISCUSSION

Our results support a staged interpretation. First, response-length control is represented before free decoding: it survives removal of decode-time instruction access and cannot be transferred primarily by the first several generated token IDs. Second, clean-recipient patching identifies the post-user template as a dominant transferable interface and middle layers as the effective depth range. Third, within-prompt actual-length trajectories contain geometry that replicates across paraphrases and prompts. Finally, targeted intervention within the frozen geometry changes lengths of natural responses that contain no length instruction.

The evidence does not imply a single scalar "length neuron." The leading direction is highly stable, but validation retains a small number of refinements, and the relevant space is re-encoded across layers. Nor do our localization results imply exclusivity: placing the length request before the user content creates a smaller additional pathway through instruction-conditioned prompt states, and the orthogonal-complement intervention remains weakly effective. The most defensible description is a dominant, low-rank, layer-indexed causal trajectory embedded in a broader distributed computation.

Our current intervention provides directional rather than precise target control. A secondary global controller saturated at its dose limits and did not reliably hit 300/500/800-token targets, so we do not present it as a successful method. More expressive prompt-adaptive trajectory coordinates may improve calibration, but that is a separate control problem. More importantly, long-directed intervention raises the automatic repetition flag from 11.7% at the natural baseline to 21.7% at dose $+0.5$. The result is causally informative but not yet a quality-preserving generation technique.

## 8 LIMITATIONS AND BROADER IMPACT

The study is limited to one instruction-tuned model, one chat-template family, greedy decoding, and English prompts drawn from one source. Therefore, our experiments establish a mechanism in Llama-3.1-8B-Instruct rather than a universal property of LMs. The 51-prompt layer scan is a discovery analysis; its broad middle-layer conclusion is supported independently by the 120-prompt causal study, but exact layer boundaries remain uncertain. The 63-prompt geometry cohort is independent of subspace training, although it had previously been inspected with a different estimator and is not a pristine never-analyzed cohort.

Length is only one aspect of response planning. It covaries with discourse structure, detail, and task difficulty, and within-cell centering cannot remove every feature that changes together with length. Cross-replication and causal controls reduce this ambiguity but do not prove that every extracted dimension represents length alone. Automated repetition and refusal filters also do not replace blinded human evaluation of correctness, relevance, and coherence.

Activation intervention could eventually provide lightweight control without rewriting user prompts, but it can also degrade generations or be used to manipulate output properties opaquely. We view the current method primarily as a mechanistic tool. Deployment would require user-visible controls, quality monitoring, stronger dose calibration, and evaluation across languages, domains, and models.

## 9 CONCLUSION

We traced response-length control from a behavioral effect to a causally active internal representation. Length-conditioned prefill states retain control without decode-time instruction access, transfer mainly through post-user template states in middle layers, and remain influential when the response opening is fixed. A cross-replicated estimator isolates a low-rank, layer-specific geometry that reproduces on independent prompts. Intervening within that frozen geometry then shortens and lengthens unseen natural responses in a dose-responsive and control-specific manner. Together, these findings provide evidence that Llama-3.1-8B-Instruct forms a causal response-length plan before generation begins, while leaving universality and precise quality-preserving control open.

## REFERENCES

Bradley Butcher, Michael O'Keefe, and James Titchener. Precise length control in large language models. *arXiv preprint arXiv:2412.11937*, 2024.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

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

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 *International Conference on Learning Representations*, 2024.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David S. Udell, Juan J. Vazquez, Ulisse Mini, and Monte Stuart MacDiarmid. Steering language models with activation engineering. 2023. URL https://api.semanticscholar.org/CorpusID:261049449.

Wilson Wu, John X. Morris, and Lionel Levine. Do language models plan ahead for future tokens? *arXiv preprint arXiv:2404.00859*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Wei Zhang, Zhenhong Zhou, Junfeng Fang, Rongwu Xu, Kun Wang, Yuanhe Zhang, Rui Wang, Ge Zhang, Xinfeng Li, Li Sun, et al. LIFEBench: Evaluating length instruction following in large language models. *arXiv preprint arXiv:2505.16234*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

## A ADDITIONAL EXPERIMENTAL DETAILS

**Mask implementation.** Length-instruction token spans are located before generation and audited by token-ID decoding. During masked decoding, attention to exactly those positions is removed while all other prompt, post-user-template, and generated-token positions remain visible. Prefix and suffix surface strings have equal semantic content but may tokenize differently at their left boundary; each condition is audited independently.

**Complete-cell geometry quality control.** For geometry extraction, a prompt–family cell is retained only when all three length tiers are valid for both paraphrase sets. A prompt must retain at least three of four families. Truncated, empty, refusal-like, and abnormally repetitive conditions are excluded before coefficient estimation. This leaves 86 prompts in the development dataset: 48 train, 18 validation, and 20 earlier held-out development prompts. Natural no-length baselines never enter discovery or hyperparameter selection.

**Permutation tests.** Within each prompt–family–paraphrase cell, the three actual-length labels are independently permuted for A and B. Coefficients and the cross-replicated operator are then recomputed. We use 500 permutations per basis and layer. Componentwise significance is not interpreted as mechanistic rank; the operational rank is chosen independently by validation reconstruction and the one-SE rule.

## B Additional Localization Results

The post-user-template effect appears in each natural baseline-length stratum. At $h_{12}$, clean recipient spans are 151.6, 149.6, and 172.8 tokens for the short, medium, and long strata, respectively, while user-position patching remains weak. In the fine scan, all three strata peak near $h_{10}$–$h_{11}$. Excluding prompts with any truncated relevant condition, retaining all prompts, or filtering pairs depth by depth yields the same qualitative ordering.

The full prefix/suffix masking experiment also constrains the localization claim. Prefix masking gives a 100-to-800 span of 463.9 tokens, versus 384.4 for suffix masking; the paired difference is 79.5 [61.4, 97.6] tokens. Normalized to the visible reference, the difference is 9.8% [7.5, 12.0]. We therefore describe the post-user template as a dominant terminal pathway, while allowing a smaller contribution from prompt states contextualized after a prefix length instruction.

## C Quality Audit for Natural-Prompt Intervention

Across the four primary $h_8$–$h_{15}$ dose conditions, truncation is 0.63% and refusal-like output is 0.83%. Abnormal repetition is 11.7% for natural baselines, 10.0% at dose $-0.5$, and 21.7% at dose $+0.5$. No development dose satisfied every preregistered safety threshold; dose 0.5 was selected by the frozen fallback rule. We therefore separate mechanistic causal efficacy from practical generation quality throughout the paper.