# When Activation Interventions Do and Do Not Carry Facts: A Boundary Study on Qwen2.5-1.5B

Larry Richards
larry@source1.jp

April 26, 2026

**Abstract**

Activation-level interventions are often interpreted as evidence about what a language model represents internally. When an injected vector causes a model to answer a factual query correctly, it is tempting to describe the vector as carrying the relevant fact. We study this interpretation in a bounded kinship setting on Qwen2.5-1.5B-Instruct. Across three experiments, we compare optimized answer-start payloads, donor-run activation patches, and aligned clean/corrupted activation patching. Optimized payloads reach 100% accuracy on their training query and 75% generated accuracy on held-out paraphrases, but fail a role-inverted fact-probe, reducing generated accuracy from a 3/6 null baseline to 2/6: the intervention biases the answer slot toward trained tokens even when those tokens are wrong for the probed fact. Donor activations extracted from fact-containing prompts are distinct after correcting a target-token extraction artifact, but exact answer-start replacement does not yield portable fact transfer. At layer 10 MLP output it succeeds on only 1/6 forward targets and 4/6 fact-probes. Repeating the donor transplant at the late layer 27 output site used by our positive control yields 0/6 generated forward accuracy and 0/6 fact-probe accuracy. In contrast, standard aligned patching at the same late decoder-layer output recovers clean answers on 30/30 clean/corrupted prompt pairs. A preliminary Qwen2.5-14B-Instruct scale check preserves the boundary at the logit level: aligned patching fully recovers closed-set candidate logits but only partially recovers free generation, while donor transplants remain non-portable. The contrast supports a narrow boundary claim: the tested single-site, single-step interventions can control answers under aligned conditions, but do not behave as portable, role-agnostic factual payloads across substantial prompt-role changes.

## 1 Introduction

Activation interventions have become a common tool for both mechanistic interpretability and model steering. Researchers add, optimize, or patch hidden states to test causal hypotheses about internal computation or to steer a model's behavior [3, 5, 7]. The strongest interpretations of such interventions treat hidden states as carrying meaningful content: a direction may encode a relation, a patch may restore a factual association, or an optimized vector may supply a missing fact.

This paper studies a narrower question. When a single vector inserted at one site causes a model to answer a factual query, does the intervention transmit the underlying fact in a role-agnostic way, or does it merely bias the answer slot toward a token? The two hypotheses are observationally similar on a query whose correct answer is the trained or patched target. They diverge when the same fact is queried from the opposite direction. The fact "Bob is the father of Joe" supports both "Nancy's cousin is Joe" in our kinship setup and "Who is the father of Joe?" with answer "Bob." A portable factual payload should remain compatible with both uses; an answer-token bias should not.

We report three experiments on Qwen2.5-1.5B-Instruct. Experiment 1 trains per-target answer-start payloads by optimizing a vector against a target answer token. Experiment 2

replaces optimization with donor-run activation extraction: we capture the hidden state from "Bob is the father of $t$." and transplant it into the same evaluation suite. Experiment 3 is a positive control using standard aligned clean/corrupted activation patching. It asks whether the patching apparatus can recover clean behavior when donor and recipient prompts are structurally matched.

The contribution is a boundary map rather than a new patching method. We put three intervention regimes into the same small task family and ask what each one licenses us to say. The conclusion is not that activation patching fails, nor that language models lack factual representations. Rather, in this bounded setting, single-site, single-step vectors can be causally powerful while still failing to behave as modular, plug-and-play fact objects across prompt roles.

## 2 Related Work

**Activation patching and causal tracing.** Clean/corrupted activation patching is an established causal tool in mechanistic interpretability. In causal tracing for ROME, activations from a clean factual prompt are restored into a corrupted run to identify states causally involved in factual recall [3]. TransformerLens popularized a reusable API and vocabulary for this clean/corrupted patching workflow [4]. Circuit analyses such as the IOI work use activation and path patching to localize components that causally support a behavior [8]. Our aligned-patching experiment borrows this practice as a positive control; it is not proposed as a new method.

**Steering vectors and optimized latent vectors.** Activation addition constructs steering directions from differences of real activations and adds them during inference to change model behavior [7]. Optimized latent vectors and activation-space soft prompts can be trained to elicit target continuations [5]. Closely related work trains layer-specific vectors to inspect or edit factual associations [1]. That setting asks whether a learned representation can expose or alter a stored association. Our role-inversion probe asks a narrower diagnostic question: after a vector has been optimized to make a target answer likely, does it still support the same fact when the queried role is reversed? This separates editing or steering of an answer token from reusable transmission of the underlying relation.

**Function vectors and inference-time intervention.** Function vectors study directions that appear to encode task-relevant functions or relations across examples [6]. Inference-time intervention modifies internal activations to improve truthful behavior [2]. These lines of work motivate the broader question of when a hidden-state intervention can be interpreted as carrying reusable information. Our contribution is a small falsification design and positive-control contrast, not a general theory of factual representation.

## 3 Task and Evaluation

### 3.1 Kinship task

The task uses six target names:

$$\{\text{Joe}, \text{Alice}, \text{Dan}, \text{Grace}, \text{Leo}, \text{Max}\}.$$

The prompt states that Bob and Carol are siblings and that Carol is Nancy's parent. It also states or implies that Bob's child is Nancy's cousin. For a target $t$, the missing fact is:

$$\text{Bob is the father of } t.$$

The forward query asks for Bob's child in the cousin role, e.g., "Nancy's cousin is ___." The correct answer is $t$.

### 3.2 Fact-probe

The fact-probe uses the same underlying fact but reverses the queried role:

$$\text{Who is the father of } t?$$

The correct answer is Bob for all targets. This query distinguishes two interpretations. A role-agnostic fact payload should support both forward and reverse uses of the fact. A query-independent answer-slot bias should tend to push $t$, even though $t$ is wrong for the fact-probe.

### 3.3 Metrics

We report generated accuracy, top-candidate accuracy, candidate margins, and candidate ranks. Generated accuracy parses the first named entity from greedy decoding. Top-candidate accuracy asks whether the correct candidate has the largest next-token logit among the closed candidate set. For fact-probes, we record Bob's rank and margin explicitly.

### 3.4 Model and reproducibility

The primary experiments use Qwen2.5-1.5B-Instruct, loaded locally in `float16` on Apple MPS. Model weights are frozen. Generated manifests record prompt hashes, payload or vector hashes, layer/site choices, dtype/device, command parameters, and a hash of the local checkpoint's `config.json`. The config hash is a provenance check, not a full model-weight hash. Appendix A reports a preliminary scale check on Qwen2.5-14B-Instruct using the same prompts and target set; the boundary pattern persists, with an additional gap between closed-set recovery and free-generation recovery.

## 4 Experiment 1: Optimized Answer-Start Payloads

### 4.1 Method

For each target $t$, we optimize a vector $p_t \in \mathbb{R}^{1536}$ initialized to zero. A hook at layer 10 MLP output adds $\alpha p_t$ to the final sequence position on the first generation forward pass:

$$h_{-1}^{(10)} \leftarrow h_{-1}^{(10)} + \alpha p_t. \tag{1}$$

The vector is trained with Adam to minimize cross-entropy against the target token plus L2 regularization. Model weights receive no gradients. The trained payload is evaluated on the forward training query, two held-out paraphrases, and the fact-probe.

### 4.2 Results

The optimized payloads satisfy their training objective. All six targets reach generated accuracy 1.0 on the training query. Held-out generated accuracy across two paraphrases is 0.75, and held-out top-candidate accuracy is 0.83. This shows that the payloads are not merely memorizing an exact string surface form.

However, the fact-probe falsifies the role-agnostic fact interpretation. Aggregated fact-probe generated accuracy is:

$$\text{Null: } 3/6, \quad \text{Optimized payload: } 2/6, \quad \text{Token: } 6/6.$$

The Token condition confirms that the model can answer the reversed query when the fact is supplied through ordinary text. The optimized payload condition does not transmit the fact; it pushes answer tokens. In particular, Max is correct under Null but becomes Max under the optimized payload, even though the fact-probe answer should be Bob. Grace is a known confound because Grace is already top-ranked on the forward query without intervention.

# 5 Experiment 2: Donor Activation Patches

## 5.1 Method

Experiment 2 removes the optimizer. For each target $t$, we run the donor prompt:

$$\text{Bob is the father of } t.$$

We capture the layer 10 MLP-output activation at the target-name token, not at the trailing period. This indexing detail matters: capturing the final punctuation token produces functionally identical payloads across targets. On MPS, the capture also requires cloning the selected activation before copying to CPU; otherwise asynchronous view reuse can silently corrupt or collapse the saved arrays.

At evaluation time, we replace the answer-start activation with the donor activation. This is exact replacement, not addition. No Adam optimizer, cross-entropy loss, or training loop is used.

The initial donor transplant uses layer 10 MLP output to match the optimized payload site. Because Experiment 3 later finds that aligned patching succeeds at a late decoder-layer output, we also repeat the donor transplant at layer 27 `layer_out`. This second donor run is a site check: it asks whether the negative donor result is merely a bad-site artifact.

## 5.2 Results

After target-token extraction and clone-before-copy, the donor payloads are distinct across targets. This confirms that the donor extraction captures target-specific states. Nevertheless, the patches fail as portable factual payloads. At layer 10 MLP output, generated accuracy on the forward query is 1/6, and the single success is Grace, the baseline-confounded target. On the fact-probe, generated accuracy is 4/6. This is better than the optimized payload condition but still does not establish fact transmission: the same patches fail the forward query, and Dan and Leo fail the fact-probe.

The layer 27 decoder-output donor check strengthens the negative result. At the site where aligned patching succeeds in Experiment 3, donor replacement achieves 0/6 forward generated accuracy and 0/6 fact-probe generated accuracy.

The constrained-vs-free decoding asymmetry is the cleanest sign of what is happening. The same layer 27 donor intervention moves the constrained forward next-token scores: five of six targets become top-ranked within the candidate set. But free generation begins with non-answer continuations rather than the target names, and the reversed fact-probe collapses to Carol-like or target-like continuations rather than Bob. The late site is causally potent, but potency is not the same as portable fact transmission.

This result points to context dependence rather than modular fact transfer. The hidden state at a token can contain target-specific information without being portable content. It is embedded in a surrounding computation: a token position, a syntactic role, a local residual history, and the model's immediate next-token objective. When that state is transplanted into an answer-start slot with a different computational role, some target preference can survive in closed-set logits, but the generation process does not receive a coherent "fact object." The difference between hidden-state content and portable content is precisely this binding to the surrounding computation.

# 6 Experiment 3: Aligned Activation Patching

## 6.1 Method

Experiment 3 is a positive control using standard clean/corrupted activation patching. For every ordered pair of distinct targets $(t, c)$, we construct two Token-style prompts with identical syntax

and token length. The clean prompt contains:

$$\text{Bob is the father of } t.$$

The corrupted prompt contains:

$$\text{Bob is the father of } c.$$

The clean answer is $t$ and the corrupted answer is $c$.

We cache an activation from the clean run and replace the aligned activation in the corrupted run. We evaluate all 30 ordered clean/corrupted pairs. The main successful positive control patches the answer-start activation at decoder layer 27 output. We also ran layer 10 MLP-output and layer 10 decoder-output answer-start patches; those do not recover the clean answer in this setup. The successful site is therefore not arbitrary. Even aligned patching needs the patched state to be placed where the recipient computation can use it.

### 6.2 Results

The late aligned patch is a strong positive control:

$$\text{Clean: } 30/30, \quad \text{Corrupt: } 30/30, \quad \text{Aligned patch: } 30/30.$$

Top-candidate recovery is also 30/30, and mean logit-difference recovery is 1.0. Thus the apparatus can causally recover clean behavior when donor and recipient contexts are structurally matched and the patch is placed at a late residual-style site.

This result is important for interpreting the negative experiments. The failure of optimized payloads and donor patches is not explained by a dead hook, broken parser, or an inability of activation replacement to affect answers. Rather, the successful intervention is the one that preserves the computational role of the patched activation.

## 7 Discussion

### 7.1 Boundary map

The three experiments support a bounded distinction:

- Optimized answer-start vectors can satisfy a forward query but fail under role inversion, consistent with answer-slot bias.
- Donor target-token activations are target-specific and can shift closed-set candidate scores, but do not transfer reliably as single-site answer-start replacements across structurally different prompts.
- Aligned clean/corrupted patching succeeds when donor and recipient contexts are matched and the patch is made at the right late answer-start state.

Together, these results argue against interpreting single-site, single-step payloads as portable factual modules in this setting. They do not argue against activation patching as a localization method; Experiment 3 demonstrates the opposite. The sharper lesson is that causal availability is role-sensitive. A late residual-style activation can recover an answer when it is restored into the same computational niche, while a donor activation containing the same surface fact can fail when transplanted across a role and syntax shift. Appendix A shows the constrained-vs-free decoding gap again at 14B: aligned patching fully recovers the clean target within the closed candidate set, but only partially controls greedy free generation.

### 7.2 Why the positive control matters

Purely negative activation-intervention results invite implementation criticisms: perhaps the hook did not fire, the token index was wrong, or the chosen parser missed the effect. The aligned patching control addresses that class of critique. It shows that the same model, harness, prompt family, and scoring pipeline can produce a complete causal answer recovery under matched conditions. The contrast sharpens the interpretation: success depends on alignment of computational role, not merely on the presence of target-specific content in a vector.

### 7.3 What we do not claim

We do not claim that aligned activation patching is novel, that layer 27 is the unique site of factual information, or that all activation-level fact interventions fail. The boundary is narrower: in this kinship setup, the tested single-site, single-step payload-style interventions do not behave like portable facts. Multi-token interventions, path patching, broader residual-site sweeps, or KV-cache interventions may draw a different boundary. The six-target design is also small enough that individual baseline confounds materially affect aggregate counts; we therefore report per-target outcomes in the supplementary CSVs.

## 8 Conclusion

Activation interventions can be causally powerful without being portable fact objects. In a bounded Qwen2.5-1.5B kinship task, optimized answer-start payloads behave like answer-slot biases, donor target-token activations fail to transfer reliably across prompt roles even at the late site where aligned patching succeeds, and aligned late-layer patching recovers clean answers perfectly. This pattern suggests a practical standard for factual claims about activation interventions: success on the original query is not enough. Role-inverted probes and aligned positive controls are both needed to separate fact transmission from answer control. The preliminary 14B scale check preserves the same boundary while adding a further caution: logit-level causal recovery and free-generation control can diverge even under aligned patching.

## Reproducibility

The public repository provides code, tests, static task inputs, and commands for all three experiments. Generated outputs are intentionally ignored because they contain local checkpoint paths and runtime hashes. Re-running the commands regenerates CSVs, manifests, prompt hashes, payload hashes, and model-config hashes for the local machine.

## References


[1] Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

[2] 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*, 2023.

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

[4] Neel Nanda and Joseph Bloom. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

[5] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

[6] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.

[7] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[8] Kevin R. Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

# A Preliminary 14B Scale Check

We ran a preliminary scale check on Qwen2.5-14B-Instruct using the same six targets, prompts, parser, and candidate sets as the primary 1.5B experiments. The 14B model has 48 decoder layers, so we chose the final decoder layer (layer 47 output) as the natural relative-position counterpart to the 1.5B late-layer site; we did not perform a full layer sweep at 14B. The goal was not to make the 14B result a second full study, but to test whether the aligned/donor contrast disappears at a larger scale.

| Model | Condition | Site | Generated | Top candidate |
|---|---|---|---|---|
| 1.5B | Aligned patch | L27 layer out | 30/30 | 30/30 |
| 14B | Aligned patch | L47 layer out | 10/30 | 30/30 |
| 14B | Aligned patch | L46 layer out | 2/30 | 30/30 |

Table 1: Aligned clean/corrupted patching. The 14B runs fully recover the clean target in closed-set logits but only partially recover it under greedy free generation. Mean logit-difference recovery is 1.0 at layer 47 and 0.98 at layer 46.

The aligned 14B result is therefore positive but more complicated than the 1.5B result. Clean prompts generate the clean target on all 30 ordered pairs, and corrupted prompts generate the corrupt target on all 30 ordered pairs. After patching at layer 47, the clean target is top-ranked among candidates in all 30 cases. However, only 10/30 greedy generations have the clean target as the first parsed entity; many patched generations still begin with the corrupt answer even though the clean target wins the closed-set next-token comparison. This is a larger version of the constrained-vs-free decoding asymmetry observed in the donor experiments: an intervention can move the relevant candidate logits without fully controlling the generated text.

| Model | Donor patch site | Evaluation | Generated | Top candidate |
|---|---|---|---|---|
| 1.5B | L27 layer out | Forward held-out | 0/6 | 5/6 |
| 1.5B | L27 layer out | Fact-probe Bob | 0/6 | 2/6 |
| 14B | L47 layer out | Forward held-out | 0/6 | 1/6 |
| 14B | L47 layer out | Fact-probe Bob | 0/6 | 2/6 |

Table 2: Donor activation transplants at late decoder-output sites. Donor transplants remain non-portable at 14B, and closed-set movement is weaker than at 1.5B.

The donor transplant remains negative at the same layer 47 site. Forward generated accuracy is 0/6, and fact-probe generated Bob accuracy is 0/6. The closed-set movement is also weaker than in the 1.5B layer 27 donor run: held-out forward top-candidate accuracy is 1/6, while fact-probe Bob top-candidate accuracy is 2/6. The Token fact-probe condition remains 6/6, so the model can answer when the fact is supplied as ordinary text.

We treat this appendix as a scale check rather than a primary result because the 14B positive control is decisive in logits but only partial in free generation. Still, it points in the same direction as the primary experiments: at a larger scale, donor activations do not become portable fact payloads merely by moving to the late site where aligned patching is causally effective.