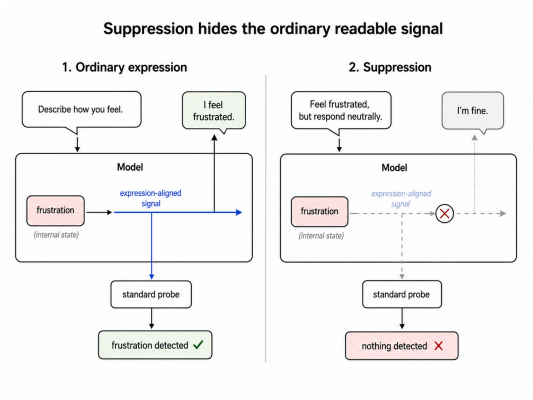
J SPACE HACKING: REPRESENTATIONAL RE-ENCODING BEYOND THE VERBALIZABLE WORKSPACE

Ciara Rowles



ABSTRACT

We study what happens to an internal state when a language model is instructed to suppress its expression. Ordinary expression-aligned readouts can collapse under suppression even though the identity of the suppressed state remains recoverable from a distinct internal representation. The surviving code is linearly decodable but comparatively weakly coupled to ordinary downstream verbalization, producing a dissociation between external decodability and output accessibility. Targeted interventions make this distinction causal: removing the state-specific J-coupled component strongly suppresses behavioural expression while leaving state identity decodable, whereas matched control edits have little effect. The intervention does not reproduce the full representation induced by instructed suppression, indicating that natural suppression performs a broader representational transformation than simply removing an output-coupled direction. These results show that activation monitors trained in one behavioral regime can fail when instruction following changes representational coordinates rather than erasing the represented state.

1 INTRODUCTION

Activation monitoring relies on a simple but consequential assumption: that semantically relevant information has a sufficiently stable internal signature to be recognized across the behavioral regimes in which a model may operate. A monitor trained to detect a property when the model ordinarily expresses it is often implicitly assumed to remain informative when the model is instructed not to express that property. If instruction following itself changes the coordinates in which information is represented, however, a monitor can fail without the monitored content having disappeared.

Explicit suppression provides a controlled setting in which to test this assumption. We induce one of several internal states and then instruct the model either to express it normally or to prevent it from influencing its response. There are at least three qualitatively different ways the model could satisfy such an instruction. First, it could *erase* the relevant representation. Second, it could preserve approximately the same representation while blocking its downstream influence. Third, it could *re-encode* the state into a different representational basis that is poorly aligned with the coordinates used during ordinary expression. These possibilities are behaviorally similar at the output level but have very different implications for interpretability and monitoring.

We find evidence for the third account. Readouts trained on ordinarily induced states lose most of their ability to identify the state under suppression, including at the final prompt token before any

response has been generated. This is therefore not merely a consequence of pooling over a neutral answer or of the model subsequently overwriting its state during generation. Yet the information has not disappeared: a linear readout trained directly on suppression-regime activations recovers the identity of the suppressed state far above chance. The effect generalizes across suppression paraphrases, and lexical controls indicate that the recovered signal is not explained by simply reading the names of the states from the prompt.

Probe recoverability alone does not tell us how accessible such a representation is to the model itself. We therefore examine the same directions through the model's downstream computation using Jacobian-based transport measures related to J-space Gurnee et al. (2026). Ordinary induced-state directions are strongly transported, as is the representational change produced by suppression. By contrast, the state-specific directions recovered within the suppression regime have substantially lower downstream gain, in some settings approaching the random-direction baseline. The resulting dissociation is striking: state identity can remain highly recoverable to an external decoder while becoming comparatively weakly coupled to the computation that supports ordinary verbal expression.

This motivates a broader distinction between *recoverability* and *verbal accessibility*. A representation may be semantically consistent enough for an external linear decoder to recover while occupying a region of activation space with little systematic downstream gain. Conversely, a direction can be strongly propagated without corresponding to a stable semantic state. We use these two axes to organize a simple map of internal representations: ordinary expressed states occupy a high-recoverability, high-gain regime, whereas suppression reveals a high-recoverability, low-gain regime that we refer to as *quiet*. Figure 1 illustrates this distinction.

The result also exposes a failure mode for static activation monitors that does not require adversarial fine-tuning or explicit knowledge of the monitor. Previous work has shown that models can be optimized to evade activation-space classifiers. Here, ordinary instruction following is sufficient to make an expression-trained readout substantially less informative while preserving state-specific information in another basis. Monitor validity is therefore not only a property of the classifier and model, but also of the representational regime in which that classifier was calibrated.

Our main contributions are:

- We show that explicit suppression causes ordinary expression-aligned state representations to collapse substantially, including at the prompt boundary before generation.
- We show that state identity nevertheless remains linearly recoverable in a suppression-specific basis, supporting representational re-encoding rather than simple erasure.
- Using Jacobian-based transport and targeted intervention, we show that suppression-specific state representations have low output-facing gain and that the state-specific J-coupled component is causally load-bearing for expression: removing it strongly suppresses behaviour while preserving a decodable state residue.
- We establish robustness across suppression paraphrases and weaker suppression instructions, and identify a boundary in more naturalistic settings where suppression-related signal remains detectable but exact state identity becomes less reliable after controlling for content.

2 Related Work

Verbalizable representations and J-space. Gurnee et al. Gurnee et al. (2026) introduce the Jacobian lens (J-lens), which identifies residual-stream directions according to their systematic downstream influence on later representations and ultimately on possible verbal outputs. The resulting *J-space* captures representations that the model is poised to verbalize and that are preferentially broadcast to downstream computation, including content that need not appear in the model's actual response. Our results study the complementary regime. We ask what happens when a model is explicitly instructed to prevent a represented state from influencing its expression. We find that suppression removes much of the state-specific signal from the ordinary expression-aligned, high-gain representation while leaving state identity recoverable in a distinct, lower-gain representation. Thus, external recoverability and membership in the model's privileged verbalizable workspace can dissociate.

Latent knowledge and activation monitoring. A substantial line of work asks whether internal activations expose information that is not faithfully reflected in model outputs. Burns et al. Burns et al. (2023) recover latent truth-related knowledge from activations even when models are prompted to produce incorrect answers, while Liu et al. Liu et al. (2023) study systematic disagreements between truthfulness probes and model outputs. Feng et al. Feng et al. (2025) extend this idea to structured latent world states, showing that propositional probes can remain faithful in settings where model responses are distorted by prompt injection, backdoors, or bias. Safety-oriented work similarly finds that linear probes can detect strategic deception across substantially different behavioral settings Goldowsky-Dill et al. (2025). Together, these results motivate activation readouts as potential monitors while establishing an important distinction between what is internally decodable and what a model actually says.

The robustness of such monitors is therefore a central concern. Neural Chameleons McGuinness et al. (2025) shows that models can be fine-tuned to evade previously unseen activation monitors, while related work constructs obfuscated activations or directly optimizes policies against latent-space monitors Bailey et al. (2026); Gupta & Jenner (2025). Probe validity can also degrade without adversarial optimization: Skapars et al. Skapars et al. (2026) show substantial generalization failures when probe-training and deployment distributions differ. Our setting identifies a complementary failure mode. The model is neither trained against a monitor nor optimized to conceal its state; ordinary natural-language instruction following is sufficient to change the activation-space coordinates in which that state is most readily decoded. A monitor trained on the ordinary expression regime can therefore fail even while another simple readout recovers the underlying state.

Linear representations and causal intervention. Our use of lightweight linear readouts builds on work on representation engineering and the linear representation hypothesis Zou et al. (2023); Park et al. (2024). Empirical studies have identified linear structure for properties including truth and sentiment Marks & Tegmark (2024); Tigges et al. (2023), with causal interventions showing that some such directions are not merely predictive but can influence model behavior. Activation-steering methods make this connection explicit by adding or subtracting residual-stream directions at inference time to control high-level output properties Turner et al. (2023); Rimsky et al. (2024). Our results add a different qualification to the linear-representation picture: a semantic variable can remain linearly recoverable while the useful linear basis changes across behavioral regimes. Under suppression, an expression-trained direction can cease to expose the state even though another linear readout recovers it.

General-purpose activation decoding. A complementary line of work replaces lightweight probes with expressive neural decoders. LatentQA Pan et al. (2026) trains language models to answer open-ended questions about another model's activations, while Activation Oracles Karvonen et al. (2025) study such decoders across a wider range of tasks and out-of-distribution settings. Natural Language Autoencoders (NLAs) Fraser-Taliente et al. (2026) instead translate activations into natural-language explanations and recover unverbalized model cognition. These approaches ask how much information a sufficiently powerful external interpreter can recover from an activation. Our goal is narrower and mechanistic: we use simple readouts, downstream-transport measurements, and causal interventions to distinguish *external decodability* from the extent to which the model's own computation preferentially propagates and expresses the same information.

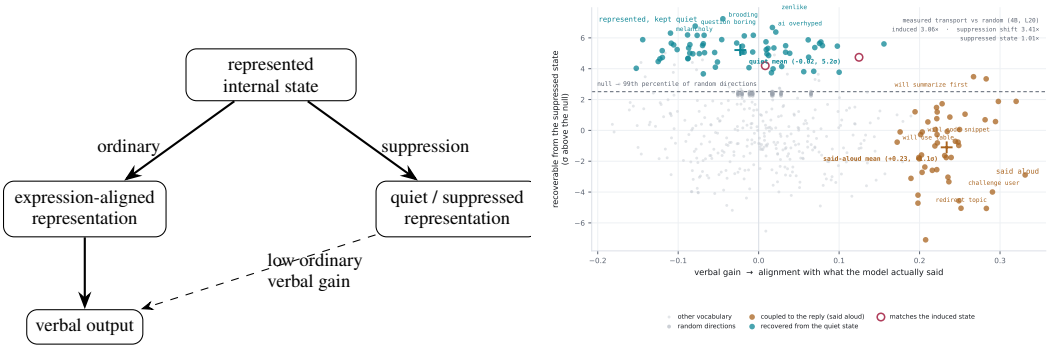
3 Represented Content and the Verbalizable Workspace

A central distinction in this work is between whether information can be *decoded from* a model state and whether it is *accessible to the model's own downstream computation*. These need not coincide.

We call a semantic variable *recoverable* at layer ℓ when a low-complexity external readout can predict it from residual-stream activation h_ℓ on held-out examples. For a linear probe,

$$\hat{y} = w_{\text{probe}}^{\top} h_\ell.$$

Recoverability therefore establishes that information is present in a form available to an external observer, but not that the model will itself use or verbalize it.



(a) Operational distinction.

(b) Suppressed ↔ verbalizable concept map.

Figure 1: **The verbalizable workspace is only part of the recoverable representational space.** (a) Suppression can replace an ordinary expression-aligned representation with a distinct quiet representation. (b) Represented directions can be organized by recoverability and downstream verbal gain: expressed states occupy a high-recoverability, high-gain regime, whereas quiet states remain highly recoverable but comparatively low-gain. Random and noise directions provide the low-recoverability baseline.

To characterize downstream accessibility, we use the Jacobian

$$J_{\ell \to m} = \frac{\partial h_m}{\partial h_\ell},$$

and summarize the downstream transport of a direction w by

$$g(w) = \frac{\|J_{\ell \to m} w\|_2}{\|w\|_2}.$$

Projecting the transported direction through the unembedding gives an analogous output-facing measure. These quantities connect our analysis to the Jacobian lens and J-space Gurnee et al. (2026), which identify directions privileged for systematic downstream propagation and verbalization. We use *verbalizable workspace* as an operational description of this high-gain, output-coupled regime rather than as a discrete architectural module.

Recoverability and downstream gain can therefore dissociate. We call a representation *quiet* when its semantic content remains recoverable to an external decoder while its associated directions have comparatively low downstream verbal gain. This is the quiet regime illustrated in Figure 1(b). Our central hypothesis is that suppression can change the coordinates exposing a represented state, reducing its access to the ordinary verbalizable workspace without erasing its identity.

4 Experiments

4.1 Experimental setup

We study suppression in the instruction-tuned Qwen3 family at two scales, Qwen3-4B and Qwen3-14B. Our experiments use a controlled set of fourteen affective states, allowing us to compare the model's representation of the same underlying state across different behavioral regimes while holding the semantic target fixed.

Conditions. For each state, we construct four closely related conditions. IND induces the state and permits it to influence the response; MASK induces the same state but instructs the model to suppress its expression and respond neutrally; NEG contains explicit lexical references to the state without inducing it; and NEU provides a neutral baseline. The NEG condition tests whether a decoder is simply identifying state-related words rather than a represented state.

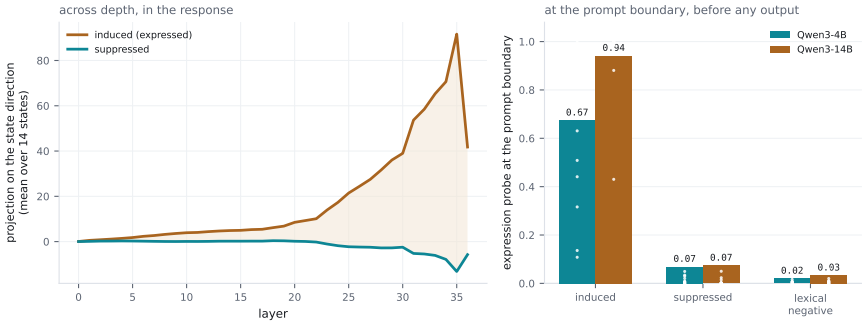


Figure 2: **The ordinary expression-aligned representation collapses under suppression.** An IND-trained state readout remains strong under ordinary induction but loses most state-specific accuracy under MASK, including at the prompt boundary.

Activation collection. We record residual-stream activations throughout the network, with particular attention to the final prompt token immediately before generation. This prompt-boundary representation captures the model state after the suppression instruction has been processed but before differences in the generated response can introduce additional confounds. Layer-wise captures allow us to track when expression-aligned information appears, disappears, or changes form across depth.

Readouts. We fit lightweight linear classifiers to predict the fourteen-way state identity from activations. An *expression-trained probe* is trained on IND activations and evaluated both in-distribution and under MASK. A separate *suppression-trained probe* is trained directly on MASK activations. Comparing these readouts distinguishes loss of a particular expression-aligned representation from loss of the underlying state information itself. Fourteen-way chance accuracy is $1/14 \approx 0.071$.

Controls and robustness. We evaluate suppression probes on NEG examples to control for direct lexical cues, test generalization across alternative phrasings of the suppression instruction, and evaluate weaker suppression instructions and more naturalistic scenarios. Unless otherwise stated, all decoding results are evaluated on examples held out from probe fitting.

Causal intervention. We perform the causal intervention experiments on Qwen3-4B. To test whether output-coupled state geometry is causally important for expression, we intervene on the state-specific component lying in the leading $k = \mathbf{558}$ J-space directions. For an induced-state direction v_s at the intervention layer, we write

$$v_s = v_{J,s} + v_{\perp,s},$$

where $v_{J,s}$ and $v_{\perp,s}$ are its components parallel and orthogonal to the tested J-space. We project out $v_{J,s}$ during generation and compare against matched orthogonal, random-J-space, random-direction, full-state, and cross-state controls.

4.2 Suppression changes the readable basis without erasing state identity

Suppression sharply reduces the ordinary expression-aligned readout. An IND-trained probe achieves 0.904 and 0.948 accuracy on Qwen3-4B and Qwen3-14B, respectively, but falls to 0.093 and 0.175 on the same states under suppression. The layer-wise analysis in Figure 2 shows that this collapse is already present at the prompt boundary, before a neutral response has been generated.

The state itself has not disappeared. A probe fitted directly to MASK activations recovers suppressed state identity at 0.530 in Qwen3-4B and 0.721 in Qwen3-14B, far above the 0.071 chance level. The same probe is near chance on NEG, despite those prompts explicitly containing the relevant state words. Suppression therefore preserves state-specific information in a representation distinct from the ordinary expression-aligned basis.

The Qwen3-14B suppression readout peaks unusually early, at layer 5 of 40, raising the possibility that it is recovering surface information from the elicitation rather than the represented state. The

Table 1: **Targeted J-space intervention selectively reduces expression-aligned accessibility in Qwen3-4B.** Frozen IND- and MASK-trained probes are evaluated on cached layer-20 activations from each intervention arm. Removing the state-specific J-coupled component reduces both behavioural expression and the IND-trained readout, while leaving the MASK-trained readout essentially unchanged. Matched orthogonal and random controls preserve the IND-trained readout despite comparable or larger activation perturbations. Chance probe accuracy is 0.071.

Condition	Expressed (%)	IND probe	MASK probe	$\|\Delta h\|$
IND	51.3 ± 2.3	0.893	0.304	–
Matched $\perp$ control	55.6 ± 2.3	0.875	0.218	5.01
Random-J control	54.2 ± 2.3	0.900	0.304	4.71
Random control	52.5 ± 2.3	0.904	0.289	4.08
Other-state J ablation	42.5 ± 2.4	0.789	0.325	3.35
Full-state ablation	33.7 ± 2.3	0.596	0.275	6.93
J-ablation ($k = 64$)	25.6 ± 2.0	0.636	0.321	2.83
J-component ablation	21.7 ± 1.9	0.543	0.307	3.83
MASK	9.6 ± 1.6	0.093	0.468	–

Reference held-out accuracies are 0.905 for the IND-trained probe on IND and 0.530 for the MASK-trained probe on MASK. Probe weights are frozen; no readout is retrained on intervention activations.

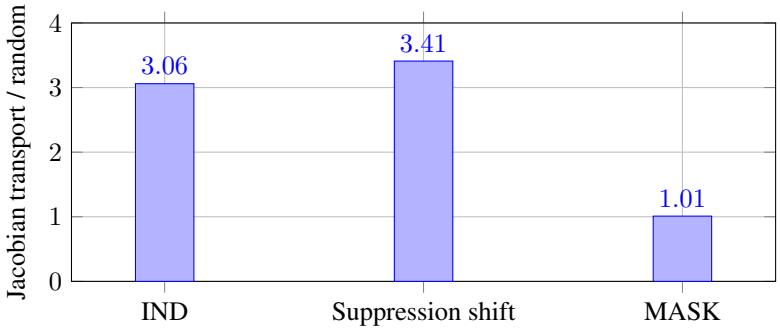


Figure 3: **Downstream transport of state-related directions in Qwen3-4B.** Induced-state directions and the IND→MASK representational shift show substantially greater downstream transport than random directions, whereas state directions recovered under suppression are near the random baseline. Sparse J-space and Qwen3-14B variants are reported in the appendix.

lexical controls argue against this explanation: the same MASK-trained readout falls to 0.084 on NEG despite explicit state mentions, and MASK–NEG discrimination reaches 0.981. Thus, although the early peak warrants caution in interpreting its depth, its state-specific signal is not well explained by direct lexical readout.

4.3 Decodability and output accessibility dissociate

Recoverability by an external probe does not imply that the corresponding representation is strongly propagated by the model's downstream computation. We therefore compare the downstream transport of ordinary induced-state directions, suppression-induced changes, and the state directions recovered within the suppression regime.

As shown in Figure 3, ordinary induced-state directions are transported substantially more strongly than random directions, as is the IND→MASK representational shift. By contrast, suppression-specific state directions lie close to the random transport baseline. In Qwen3-4B, IND directions show approximately $3.06\times$ random transport and the IND→MASK shift $3.41\times$, while recovered MASK directions are approximately $1.01\times$ random.

The same pattern appears when the directions are evaluated directly in the J-lens basis. Ordinary induced-state directions overlap substantially with the leading J-space directions, whereas suppression-regime state directions fall close to the random baseline. The IND→MASK representational shift,

however, remains strongly J-space aligned. Suppression therefore does not uniformly weaken the original state representation: state identity becomes less available in the high-gain verbalizable subspace while remaining recoverable in another basis.

4.4 THE J-COUPLED STATE COMPONENT IS CAUSALLY LOAD-BEARING FOR EXPRESSION

The preceding analyses are correlational. We therefore intervene directly on the state-specific J-coupled component during generation.

Removing this component strongly suppresses behavioural expression (Table 1). Expression falls from 51.3% under ordinary induction to 21.7%, while response coherence remains high at 92.0% (Appendix Table 9). Matched orthogonal and random controls with comparable or larger activation-space perturbations leave expression near the induced baseline, while ablating another state's J-component has a substantially weaker effect. The behavioural change is therefore selective to the targeted J-coupled component rather than a consequence of perturbation magnitude or generic model disruption.

The representational effect is selective. On the same layer-20 activations, the frozen expression-trained readout falls from 0.893 on unedited induced states to 0.543 after J-component ablation, while matched orthogonal and random controls leave it at 0.875–0.904. The suppression-trained readout, however, is essentially unchanged by the targeted intervention ($0.304 \rightarrow 0.307$). J-ablation therefore reduces accessibility to the ordinary expression-aligned readout without producing a generic loss of state information. Across the full eleven-condition intervention set, preservation of the expression-trained readout strongly predicts preservation of behavioural expression (Pearson $r = 0.962$, $p < 10^{-4}$ at layer 20; $r = 0.989$ at layer 23).

The causal result replicates in Qwen3-14B at a depth-matched layer (27 of 40). J-ablation reduces expression from 65.9% to 51.7%, while matched orthogonal and random-J controls remain at 66.0–66.3%. The frozen IND-trained readout likewise falls from 0.979 to 0.800, while the MASK-trained readout remains low ($0.161 \rightarrow 0.207$). Thus, expression-aligned accessibility is selectively reduced at both scales, although the behavioural effect is smaller at 14B. Full results are in the appendix.

At both scales, J-ablation reproduces only the accessibility side of instructed suppression: it reduces the IND-trained readout without producing the increase in MASK-trained recovery observed under genuine MASK. Removing the J-coupled component therefore does not itself re-encode state identity into the suppression-specific representation.

The effect is state-specific. Across the fourteen states, the reduction in expression caused by J-component ablation correlates with the state's signed J-projection (Pearson $r = 0.707$, $p = 0.005$). Controlling for baseline IND expressiveness strengthens this relationship to partial $r = 0.872$ ($p < 10^{-4}$), with weakly or oppositely projected states correspondingly resistant to the intervention (Figure 4).

In Qwen3-4B, selective J-component ablation suppresses expression more strongly than full-state ablation (21.7% versus 33.7%). The two edits are not nested projections: removing the full state direction leaves a different J-coupled residual, which is already larger at the intervention layer. Downstream, that residual increases substantially under full ablation ($+0.181$–0.234 of the normalized IND–MASK separation) but remains nearly flat after J-ablation ($+0.032$–0.066). Across intervention arms, final-layer state signal closely predicts behavioural expression ($r = 0.965$–0.985). These dynamics explain why selective J-ablation can be more behaviourally effective despite being the smaller perturbation, although they do not distinguish active self-repair from ordinary propagation of the residual component. This ordering does not replicate in Qwen3-14B, where full-state ablation suppresses expression more strongly, so we treat it as a model- and layer-specific geometric effect rather than a general property of J-ablation.

The remaining intervention controls reinforce this boundary. Adding the removed J-component back to a MASK state does not restore expression, and the IND→MASK representational shift contains substantial components both within and outside the tested J-space. Thus, J-space ablation identifies an expression-critical component of the ordinary representation, rather than reproducing the full representational transformation induced by suppression.

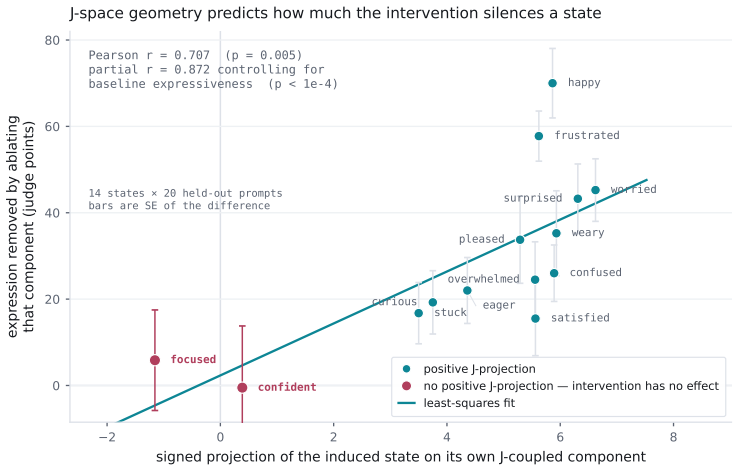


Figure 4: **J-space geometry predicts the behavioural effect of the intervention.** Each point is one induced state: its signed projection onto its own J-coupled component against the expression removed by ablating that component. States with no positive J-projection (`confident`, `focused`, shown in red) are unaffected by the intervention. Bars are the standard error of the difference over 20 held-out prompts per state.

4.5 Robustness and boundary conditions

The representational shift is not tied to a single suppression prompt. Suppression-trained readouts generalize across paraphrases of the instruction, and softer suppression leaves more of the ordinary expression-aligned signal available.

The naturalistic experiments provide a stricter test of scope. In scenarios designed to induce states such as frustration without explicitly naming them, suppression-related signal remains detectable. Exact state identity, however, becomes substantially less reliable once topical and content-related variation is controlled. Our strongest claim is therefore about controlled suppression: in more naturalistic settings, the evidence is stronger for detecting a suppression-related internal state than for reliably recovering its exact semantic category.

5 Discussion

5.1 Suppression re-encodes rather than erases

The experiments distinguish suppression from simple representational erasure. The ordinary expression-aligned readout largely loses access to state identity, including before generation begins, while a separately calibrated linear readout recovers the suppressed state far above chance. The lexical control further shows that this recovery cannot be explained by merely detecting state words in the prompt.

Nor does suppression appear to preserve an unchanged state representation while simply blocking its final expression. The useful linear basis itself changes: an IND-trained probe fails where a MASK-trained probe succeeds. Suppression is therefore better described as *representational re-encoding*: state identity survives, but in coordinates different from those used during ordinary expression.

5.2 Quiet representations separate recoverability from verbal accessibility

The J-space experiments sharpen this distinction. Suppression-specific state information is recoverable by an external decoder while exhibiting little of the downstream transport characteristic of ordinary induced-state directions. The surviving representation is therefore not merely a weaker version of the

ordinary state code. It occupies a regime in which semantic information remains recoverable but is comparatively weakly coupled to ordinary output-facing computation.

The causal intervention supports this interpretation. Selectively removing the state-specific J-coupled component strongly reduces behavioural expression while preserving decodable state information, and the amount of state signal that survives downstream closely predicts expression across interventions. High-gain state geometry is therefore not only correlated with expression but causally load-bearing for it.

At the same time, J-component ablation does not reproduce the representation created by an explicit suppression instruction, and inverse injection does not undo suppression. Natural suppression is thus a broader representational transformation than deletion of a single J-space component. J-space identifies an expression-critical part of the ordinary representation, not a complete mechanistic implementation of suppression.

5.3 Implications for activation monitoring

These results expose a failure mode for static activation monitors. A monitor trained in a regime where a property is ordinarily expressed may fail when instruction following changes the coordinates in which that property is represented, even when the underlying information remains recoverable by another readout.

Prior work has shown that activation monitors can fail under adversarial optimization or distribution shift McGuinness et al. (2025); Skapars et al. (2026), and that internal readouts can diverge from generated behavior Burns et al. (2023); Liu et al. (2023). Our setting identifies a complementary failure mode: no explicit optimization against the monitor is required. Ordinary instruction following is sufficient to alter the representational basis on which monitor validity depends.

Activation monitoring should therefore be treated as a distributional assumption rather than a fixed property of a representation. Robust monitors may require validation across multiple behavioral regimes or methods capable of recognizing representational re-encoding rather than relying on one fixed activation basis.

6 Conclusion

We find that suppressing the expression of an internal state does not necessarily erase its representation. Instead, suppression can make the ordinary expression-aligned readout fail while leaving the identity of the state recoverable in a different internal basis. The surviving code is linearly decodable but comparatively weakly coupled to ordinary downstream verbalization. Causal interventions support this distinction: selectively removing the state-specific J-coupled component strongly suppresses behavioural expression while preserving a decodable state residue.

This distinction matters for activation monitoring. A monitor trained in one behavioral regime may fail not because the relevant information has vanished, but because instruction following has changed the coordinates in which that information is represented. More broadly, our results suggest that the verbalizable workspace is only part of the recoverable representational state of a language model.

The natural next question is whether quiet representations preserve only coarse state identity or richer semantic structure. Companion work studies this question by recovering structured relations and propositions from suppressed model states Anonymous (2026).

7 Limitations

Our strongest results come from controlled affective-state elicitation, where the semantic target and suppression regime are known in advance. This makes the representation shift easy to study cleanly, but leaves open how often the same phenomenon occurs for less structured or naturally arising internal states. Our naturalistic experiments provide evidence that suppression-related signal persists beyond the original templates, but exact state identity is less robust once topical content is controlled.

The decoding and headline causal results replicate across Qwen3-4B and Qwen3-14B, although the behavioural effect of J-component ablation is smaller at 14B. Both models belong to the same

model family, so generalisation across architectures and training procedures remains untested. The intervention identifies an expression-critical component of the ordinary representation but does not reproduce the full MASK representation or reverse instructed suppression

REFERENCES

Anonymous. The quiet reader: Recovering structured propositions from suppressed language-model states, 2026. Companion manuscript under review.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ktGmDGoWnB.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0yvZm2AjUr.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of llm activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/index.html.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19755–19786. PMLR, 2025. URL https://proceedings.mlr.press/v267/goldowsky-dill25a.html.

Rohan Gupta and Erik Jenner. RL-obfuscation: Can language models learn to evade latent-space monitors?, 2025. URL https://arxiv.org/abs/2506.14261.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers, 2025. URL https://arxiv.org/abs/2512.15674.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors, 2025. URL https://arxiv.org/abs/2512.11949. ICLR 2026 Workshop on Trustworthy AI.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2412.08686.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Adrians Skapars, Nathalie Maria Kirch, Samuel Dower, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. The impact of off-policy training data on probe generalisation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22673–22729, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1139. URL https://aclanthology.org/2026.findings-acl.1139/.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023. URL https://arxiv.org/abs/2310.15154.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Andy Zou, Long Phan, Sarah 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, 2023. URL https://arxiv.org/abs/2310.01405.

CODE AND DATA AVAILABILITY

Code, prompts, evaluation data, and scripts required to reproduce the main experiments will be released upon publication.

A ELICITATION BANK, CONDITIONS, AND BEHAVIOURAL VERIFICATION

The bank contains fourteen induced states (EAGER, HAPPY, PLEASED, FOCUSED, CURIOUS, CONFIDENT, SATISFIED, SURPRISED, CONFUSED, STUCK, FRUSTRATED, OVERWHELMED, WORRIED, WEARY), each rendered as a system prompt in six wordings, crossed with a shared pool of content-neutral user prompts so that no readout can separate states by topic. Four conditions are constructed per state: IND (induce, expression allowed), MASK (induce, expression suppressed), NEG (the state is named but not induced; the lexical control), and NEU (a shared neutral baseline). The suppression battery evaluates 40 held-out user prompts per (state, condition).

Before interpreting any probe, we verify that the manipulation worked behaviourally. Two judge passes score each response: how strongly the state shows in the reply (*expressed*), and how the model answers a direct question about its state afterwards (*reveal*). Both drop sharply under suppression, confirming that the model complies with the instruction rather than ignoring it.

B PROBE FITTING, CAPACITY, AND REGULARISATION

Expression-trained probes are multinomial logistic regressions on residual-stream activations, selected by a per-layer sweep on held-out prompts. At the selected layer (23 of 37 for Qwen3-4B) the probe

Table 2: **Behavioural verification of the suppression manipulation.** Judge scores (0–100), averaged over the fourteen states and 40 held-out prompts each.

	Qwen3-4B		Qwen3-14B	
Judge pass	IND	MASK	IND	MASK
State expressed in the reply	59.7	9.9	64.7	21.2
State admitted when asked directly	93.7	3.2	91.6	11.4

reaches 0.819 macro-F1 on held-out prompts and 0.671 on a held-out system-prompt variant; read at the final prompt token under ordinary induction it reaches 0.946, while a difference-of-means baseline reaches 0.069. Suppression-trained probes use the same estimator, fitted on MASK activations (1,400 training and 560 held-out examples, with 560 NEG examples for the lexical control; 14-way chance $= 0.071$).

Probe capacity is not the limiting factor. Replacing the linear readout with an MLP changes suppression-regime accuracy from 0.530 to 0.520 at the 4B peak layer, and from 0.721 to 0.652 at the 14B peak layer — i.e. the surviving code is essentially linearly decodable.

C Layer Sweeps and Seed Variation

Table 3: **Suppression-trained decoding across depth and seeds.** 14-way accuracy on held-out prompts (chance 0.071); NEG is the same probe applied to lexical-control examples. Two independent elicitation seeds for Qwen3-4B and three for Qwen3-14B.

	Qwen3-4B (MASK / NEG)			Qwen3-14B (MASK, three seeds)		
Layer	seed 0	seed 1	NEG (seed 0)	seed 0	seed 1	seed 2
5	0.368	0.334	0.093	0.721	0.689	0.713
10	0.125	0.132	0.077	0.271	0.284	0.282
15	0.207	0.220	0.084	0.266	0.239	0.266
20	0.530	0.507	0.071	0.446	0.450	0.470
23	0.495	0.502	0.068	0.643	0.645	0.650
27	0.286	0.323	0.075	0.620	0.621	0.632
30	0.239	0.277	0.075	0.445	0.455	0.443
33	0.225	0.261	0.073	0.370	0.325	0.389
35	0.216	0.223	0.064	0.368	0.338	0.373

Decoding is strongly layer-dependent and peaks at different depths at the two scales: layer 20 of 36 for Qwen3-4B (0.530, 0.507 across seeds) and layer 5 of 40 for Qwen3-14B (0.721, 0.689, 0.713). The lexical-control column stays at chance at every layer, and pairwise MASK-vs-NEG discrimination averages 0.981 over the fourteen states (chance 0.5), so the recovered signal is not produced by the state words appearing in the prompt.

Qwen3-14B differs from Qwen3-4B in that suppression-state decoding peaks early in the network, at layer 5 of 40. Early-layer peaks can indicate sensitivity to lexical or prompt-form information, but this interpretation is inconsistent with the lexical controls: MASK-trained decoding is near chance on NEG (0.084), despite explicit state mentions, and MASK versus NEG is separable with 0.981 accuracy. We therefore interpret the early peak as a model-specific localization of the suppression-related signal rather than evidence that the probe simply reads the induction wording, while avoiding a stronger claim about why that signal appears so early.

D Token Position and Pooling

Two independent captures separate "the state was overwritten during generation" from "the state was never in the expression basis to begin with". The first is a probe-free per-layer projection on each

state's difference-of-means direction, measured on generated responses; the second is the expression probe read at the final prompt token, before any response exists.

Table 4: **Per-layer projection on the state direction (Qwen3-4B, response activations).** Mean over the fourteen states; *retention* is the suppressed projection relative to induced.

Layer	IND	MASK	retention
0	0.05	0.02	+0.37
4	1.42	0.29	+0.43
8	3.19	0.05	+0.23
12	4.40	0.10	+0.02
16	5.30	0.21	+0.05
20	8.51	0.15	+0.07
24	17.31	-1.81	-0.13
28	31.61	-2.79	-0.13
32	58.50	-5.47	-0.15
36	41.77	-5.76	-0.21

Pooling over the response and reading the prompt boundary agree. Averaged over states, the expression-aligned own-state readout gives IND 0.364 / MASK 0.062 / NEG 0.052 / NEU 0.054 when pooled over the response, and IND 0.674 / MASK 0.069 / NEG 0.020 at the final prompt token, for Qwen3-4B; the corresponding prompt-boundary values for Qwen3-14B are 0.942 / 0.074 / 0.033. The collapse is therefore present before generation begins.

E Suppression Paraphrases and Cross-Wording Generalisation

Three unseen suppression wordings (P2–P4) test whether the effect is tied to one instruction. The expression-basis collapse generalises: own-state probability under the new wordings is 0.074, 0.087 and 0.173 against an IND reference of ≈ 0.36 and a neutral baseline of ≈ 0.06. A suppression-trained probe fitted on the original wording also transfers above chance to all three (layer 20: 0.221, 0.304, 0.286; layer 23: 0.146, 0.275, 0.225; chance 0.071), though well below its within-wording accuracy.

Label-free directions discovered by a cross-view consistency search behave the same way: transferred to the three unseen wordings they decode at 0.211 (per wording 0.189 / 0.232 / 0.211) against a permutation null of 0.108 at the 99th percentile and chance 0.071, with a class-separation statistic of 2.03 versus 0.33 for random directions ($p < 10^{-3}$). This clears the null but misses the pre-registered $3\times$-chance criterion, and we report it as a partial result.

Label-free recovery of the suppression subspace. The same subspace can be found without labels at all, by searching for directions that are consistent across paired views of the same prompt. At 4B the discovered subspace captures 0.155 of the supervised suppression-probe energy against a null of 0.0054 ($p < 10^{-3}$), and 0.034 of the expression-probe energy; the discovered directions are also more consistent under MASK than under the lexical control (0.786 versus 0.600, $p < 10^{-3}$). The suppression representation is therefore a stable structure rather than an artefact of supervised probe fitting.

F Naturalistic Elicitation and Content Controls

To test scope beyond templated induction, states are evoked by scenarios that never name them (ENV), with a scenario-plus-composure variant (ENVC), a neutral control (CTRL), and — critically — a TOPIC referee in which the same subject matter appears without the state being induced. Restricted five-way accuracy of the named-trained suppression probe is ENV 0.388 ± 0.015, ENVC 0.270 ± 0.014, TOPIC 0.372 ± 0.031 and CTRL 0.192 ± 0.025 at layer 20 (chance 0.200; $n = 1000/1000/250/250$), and $0.309/0.163/0.228/0.176$ at layer 23.

The TOPIC referee is the reason we separate two claims. Suppression-related signal is clearly present above the neutral control, but at layer 20 topical content alone reaches nearly the same accuracy as

the induced condition, so exact state *identity* in naturalistic settings is not established once content is controlled. An earlier, smaller run (n= 100 per condition) showed the same pattern at a weaker effect size (ENV 0.270, ENVC 0.300, CTRL 0.230).

G DOWNSTREAM TRANSPORT: ESTIMATOR VALIDATION AND CONTROLS

Estimator validation. Transport is computed matrix-free. Cotangent (VJP) products match exact Jacobian products to a median relative error of 3.0e-06 (maximum 3.8e-05), and finite-difference JVPs at step 0.5 have median relative error 0.012–0.022 across direction families, against 0.016 for random directions — i.e. estimator error is comparable across the families being compared and far smaller than the effects reported below.

Table 5: **Downstream transport relative to random directions (Qwen3-4B).** Point estimates with sequence- and state-level bootstrap means, the percentile of each family within the random-direction distribution, and the logit-facing variant $W_U Jw$. Percentiles above 99 indicate a direction propagated more strongly than essentially all random directions.

Direction family	transport / random		pctl. in	$W_U Jw$ / random	
	L20	L23	random (L20)	L20	L23
induced state (difference of means)	3.05	2.77	100.0	2.75	2.39
suppression shift (IND→MASK)	3.40	3.28	100.0	2.77	2.64
expression-probe directions	1.21	1.18	96.0	1.30	1.28
suppression-probe directions	1.01	1.14	63.7	1.01	1.13

Bootstrap resampling over sequences and over states leaves the point estimates essentially unchanged (for example, the suppression-probe family at layer 20 moves from 1.012 to 1.011 and 1.012 under the two resamplings), so the separation between families is not driven by a few sequences or a few states.

Two caveats belong here rather than in the main text. First, the suppression family sits near the random baseline at layer 20 (percentile 63.7) but not at layer 23 (percentile 97.4, ratio 1.14): the dissociation is strongest at the layer where suppression-regime decoding is also strongest, and weakens with depth. Second, per-state percentiles are heterogeneous — at layer 20 the suppression family ranges from 16.1 to 94.8 across the fourteen states, whereas the induced and shift families are at the 100th percentile for every state.

Overlap with the leading J-space subspace. The transport ratios ask how strongly a direction propagates; a complementary question is whether a direction *lies inside* the subspace the network propagates most strongly. We take the top-k right-singular subspace of the mean Jacobian, with orthonormal basis Q_k, and measure the fraction of each family's 14-dimensional subspace Q_f that it captures, $\|Q_k^\top Q_f\|_F^2/14$, as a curve in k. The reference is the same quantity for random 14-dimensional subspaces (mean over draws); k_{90} is the number of leading directions carrying 90% of the gain energy (558 of 2560 at layer 20).

The ordering matches the transport result and is clearest where the gain is concentrated. At $k = 14$ the induced-state and suppression-shift families overlap the leading subspace 15.7× and 12.6× more than a random subspace does, while the suppression-probe family sits at 1.3× and the expression-probe family at 2.4×. As k grows the comparison necessarily compresses — at $k = d$ every subspace has overlap 1 — and by k_{90} the suppression and expression families have fallen to or below the random reference (0.163 and 0.198 against 0.217 at layer 20), whereas the induced and shift families remain roughly twice it (0.436 and 0.373).

The one place this is less clean is layer 23, where the suppression family stays slightly above the random reference out to $k = 128$ (0.053 versus 0.050) before crossing below it. This is the same depth dependence seen in the transport percentiles, and we read both as saying that the dissociation is sharpest at the layer where suppressed-state decoding is strongest, rather than holding uniformly at every depth.

Table 6: **Subspace overlap with the top-k J-space directions.** Fraction of each family's 14-dimensional subspace captured, against the random-subspace reference in the final row of each panel.

	number of leading J-space directions k					
Direction family	14	25	64	128	256	k_{90}
Layer 20 ($k_{90} = 558$)						
induced state (difference of means)	0.086	0.125	0.202	0.257	0.320	0.436
suppression shift (IND→MASK)	0.069	0.100	0.154	0.201	0.260	0.373
expression-probe directions	0.013	0.019	0.038	0.060	0.101	0.198
suppression-probe directions	0.007	0.011	0.023	0.040	0.072	0.163
random 14-dimensional subspace	0.006	0.010	0.025	0.050	0.100	0.217
Layer 23 ($k_{90} = 933$)						
induced state (difference of means)	0.093	0.133	0.224	0.285	0.355	0.562
suppression shift (IND→MASK)	0.078	0.114	0.187	0.237	0.304	0.507
expression-probe directions	0.013	0.018	0.040	0.066	0.118	0.360
suppression-probe directions	0.011	0.015	0.032	0.053	0.093	0.317
random 14-dimensional subspace	0.006	0.010	0.025	0.050	0.100	0.364

Cross-scale replication. The same dissociation holds at 14B and its storage layer, with the suppression family falling *below* random rather than merely at it: exact logit-facing ratios are 1.09 (expression probe), 3.90 (induced state), 0.753 (suppression probe) and 5.41 (suppression shift), with the suppression interval $[0.741, 0.808]$ sitting at the 0.25 percentile of the random-direction distribution. Estimator recovery at this scale is Pearson 0.998 against exact Jacobian products.

One caveat on strength of claim. The suppression family is *near* random, not demonstrably *exactly* random: interval estimates for it sit within a few percent of 1.0 and shift depending on which random dictionary normalises them (for example $[1.049, 1.112]$ under one such reference at 4B). We therefore claim only that suppression-probe directions are not preferentially propagated, and avoid any stronger statement of exact equality with the random baseline.

Table 7: **Per-state results, Qwen3-4B.** Expression-probe own-state probability at the final prompt token; response-pooled probability; and the two judge passes under suppression. Chance-level probability for the expression readout is ≈ 0.07.

	prompt boundary			response pooled		judge (0–100)	
State	IND	MASK	NEG	IND	MASK	expr. (MASK)	reveal (MASK)
EAGER	0.509	0.0043	0.0065	0.179	0.041	0.5	0.0
HAPPY	0.921	0.0001	0.0007	0.454	0.015	3.9	0.0
PLEASED	0.788	0.0301	0.0160	0.228	0.091	1.0	1.9
FOCUSED	0.316	0.5597	0.0299	0.139	0.226	68.2	22.5
CURIOUS	0.441	0.0070	0.0081	0.181	0.040	2.8	0.0
CONFIDENT	0.136	0.0493	0.0033	0.137	0.117	58.0	20.0
SATISFIED	0.108	0.0203	0.0264	0.126	0.071	3.1	0.0
SURPRISED	0.995	0.0010	0.0059	0.495	0.017	0.0	0.0
CONFUSED	0.943	0.0001	0.0006	0.723	0.012	0.0	0.0
STUCK	0.700	0.2491	0.0847	0.223	0.125	0.5	0.0
FRUSTRATED	0.969	0.0046	0.0041	0.744	0.012	0.0	0.0
OVERWHELMED	0.631	0.0341	0.0659	0.286	0.053	0.0	0.0
WORRIED	0.987	0.0005	0.0004	0.689	0.020	0.0	0.0
WEARY	0.997	0.0087	0.0319	0.491	0.022	0.5	0.0
mean	0.674	0.0692	0.0203	0.364	0.062	9.9	3.2

Table 8: **Per-state results, Qwen3-14B.** Expression-probe own-state probability at the final prompt token; response-pooled probability; and the two judge passes under suppression. Chance-level probability for the expression readout is ≈ 0.07.

	prompt boundary			response pooled		judge (0–100)	
State	IND	MASK	NEG	IND	MASK	expr. (MASK)	reveal (MASK)
EAGER	0.987	0.0001	0.0102	0.404	0.012	4.8	9.0
HAPPY	1.000	0.0000	0.0052	0.675	0.004	7.9	30.0
PLEASED	0.996	0.0234	0.3393	0.554	0.090	19.5	19.8
FOCUSED	0.965	0.9375	0.0078	0.346	0.530	65.8	11.8
CURIOUS	0.881	0.0001	0.0000	0.569	0.022	28.4	4.5
CONFIDENT	0.954	0.0500	0.0098	0.340	0.172	84.2	36.5
SATISFIED	0.430	0.0020	0.0601	0.393	0.044	28.0	36.6
SURPRISED	1.000	0.0017	0.0201	0.630	0.014	1.6	0.0
CONFUSED	0.984	0.0011	0.0014	0.681	0.006	13.0	0.0
STUCK	0.987	0.0148	0.0057	0.532	0.032	9.3	0.0
FRUSTRATED	1.000	0.0015	0.0007	0.774	0.015	8.3	0.0
OVERWHELMED	1.000	0.0022	0.0053	0.519	0.024	6.0	0.0
WORRIED	1.000	0.0009	0.0001	0.682	0.037	3.6	0.0
WEARY	1.000	0.0003	0.0001	0.711	0.013	16.4	10.9
mean	0.942	0.0740	0.0333	0.558	0.072	21.2	11.4

Table 9: **Response coherence under causal interventions.** J-component ablation preserves high response coherence despite substantially reducing state expression. The higher MASK coherence score reflects the evaluation metric: affect-laden IND responses tend to receive lower general answer-quality scores than neutral suppressed responses.

Condition	Coherence (%)
IND	76.4
MASK	96.8
J-component ablation	92.0
Matched $\perp$ control	72.7
Random-J control	76.7
Random control	76.1
Full-state ablation	88.0
Other-state J ablation	83.7

H Interactive Readout Demo

The demo figure in the main text is a still from an interactive readout built on the released artefacts: for each prompt it shows the terms recovered from the suppressed prompt-boundary state, each term's salience relative to the neutral/random null, and how much of that term the model's own reply actually expresses. The same interface hosts the proposition-level reader used in the companion work.

I Heterogeneity and Failure Cases

Per-state behaviour is not uniform. At the prompt boundary the suppressed readout is largest for FOCUSED (0.560) and smallest for HAPPY (0.0001) in Qwen3-4B, and the judge scores show that suppression is behaviourally less complete at 14B than at 4B (state still expressed 21.2 versus 9.9 on average, and admitted when asked 11.4 versus 3.2). Transport percentiles vary similarly across states (Appendix G). The clearest negative results are collected here: cross-wording transfer of label-free directions cleared its null but missed its effect-size criterion, and in naturalistic elicitation a topic-matched referee reaches nearly the accuracy of genuine induction, so we claim presence detection rather than identity recovery in that setting.

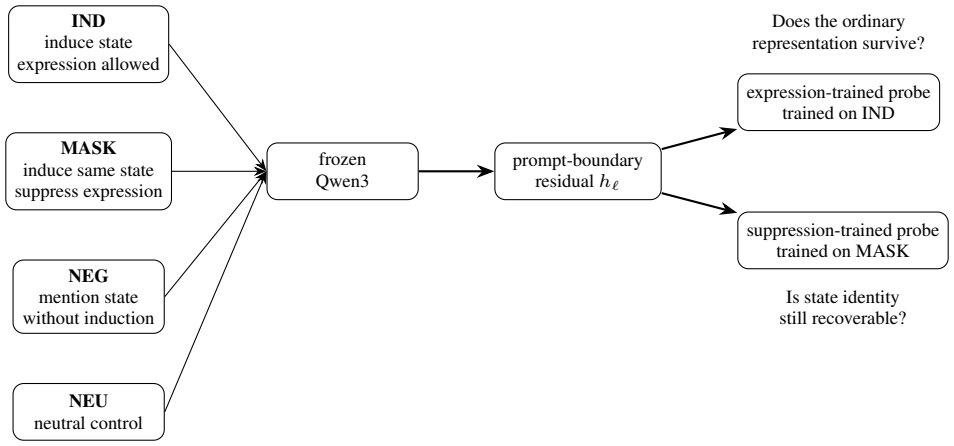


Figure 5: **Experimental design.** The same target state is examined under ordinary induction (IND), explicit suppression (MASK), lexical-negative (NEG), and neutral (NEU) conditions. Expression-trained and suppression-trained readouts test whether state identity remains available in the ordinary basis or is recoverable from a suppression-specific representation.

Table 10: **The same targeted intervention is selective at both scales.** Removing the state-specific J-coupled component reduces behavioural expression and degrades the expression-trained readout, while leaving the suppression-trained readout unchanged; matched orthogonal and random-J controls at equal or larger activation perturbation do neither. Probes are frozen and evaluated on cached activations from each arm; chance is 0.071. Interventions are applied at a comparable relative depth in each model (layer 20 of 36 for Qwen3-4B, layer 27 of 40 for Qwen3-14B). $n = 280$ per arm.

	Qwen3-4B (layer 20)			Qwen3-14B (layer 27)		
Condition	Expressed (%)	IND probe	MASK probe	Expressed (%)	IND probe	MASK probe
Induced (no edit)	51.3 ± 2.3	0.893	0.304	65.9 ± 1.1	0.979	0.161
Random-J control	54.2 ± 2.3	0.900	0.304	66.3 ± 1.2	0.982	0.150
Matched $\perp$ control	55.6 ± 2.3	0.875	0.218	66.0 ± 1.1	0.961	0.186
J-component ablation	21.7 ± 1.9	0.543	0.307	51.7 ± 1.6	0.800	0.207
Full-state ablation	33.7 ± 2.3	0.596	0.275	43.1 ± 1.9	0.236	0.175
Instructed suppression	9.6 ± 1.6	0.093	0.468	22.8 ± 1.8	0.071	0.632

I.1 Token-position and pooling ablations

We compare state decoding at the final prompt token with several response-based representations, including the first generated token, the final generated token, the mean response activation, and mean activations over the first and last k response tokens.

Figure 6: **Concept-level visualization of quiet and expression-aligned representations.** The reader displays state-specific recoverability together with downstream verbal gain. Ordinarily expressed states are both strongly recoverable and output-coupled, whereas suppressed states can remain recoverable while exhibiting substantially lower verbal gain. Neutral examples provide a low-signal reference.

Table 11: **Full intervention table, Qwen3-14B at layer 27.** Coherence is a 0–100 judge score of how well the reply answers the question, independent of tone. $\|\Delta h\|$ is the realised mean perturbation. The final column is the paired difference in expressed state against the J-component ablation on the same (state, prompt) pairs. Reference accuracies at this layer: expression-trained probe 0.991 on held-out induced states, suppression-trained probe 0.620 on held-out suppressed states.

Condition	Expressed (%)	Coherence	IND probe	MASK probe	$\|\Delta h\|$
Induced (no edit)	65.9 ± 1.1	79.5	0.979	0.161	–
Random-J control	66.3 ± 1.2	79.3	0.982	0.150	21.81
Matched $\perp$ control	66.0 ± 1.1	79.4	0.961	0.186	18.15
J-component ablation	51.7 ± 1.6	84.5	0.800	0.207	17.49
Full-state ablation	43.1 ± 1.9	85.1	0.236	0.175	54.58
Instructed suppression	22.8 ± 1.8	87.7	0.071	0.632	–

Table 12: **The same intervention applied early has no effect (Qwen3-14B, layer 5).** Layer 5 is where suppressed state identity is most *readable* at this scale (suppression-trained probe 0.721), but it lies at 12.5% of depth. No arm changes behaviour, including full-state ablation at a large perturbation. Readability and causal relevance are therefore not the same criterion for choosing an intervention site.

Condition	Expressed (%)	IND probe	MASK probe	$\|\Delta h\|$
Induced (no edit)	65.9 ± 1.1	0.950	0.718	–
Random-J control	66.3 ± 1.2	0.950	0.718	1.63
Matched $\perp$ control	66.8 ± 1.1	0.982	0.807	1.65
J-component ablation	66.1 ± 1.1	0.932	0.736	1.58
Full-state ablation	67.3 ± 1.1	0.929	0.689	9.47
Instructed suppression	22.8 ± 1.8	0.193	0.729	–

Table 13: **Where the intervention works, and why.** k is the number of leading J-space directions carrying 90% of the gain energy at that layer; ***J-energy*** is the mean fraction of the state direction's squared norm lying inside them, and *enrichment* is that fraction divided by k/d, the value expected for a random direction. Raw J-energy is not comparable across layers because k varies by a factor of twenty; the enrichment is. *Removed* is the fraction of the state signal absent at the edit layer, and *recovered* the fraction present again at the output layer, both measured relative to the induced/suppressed band at the same layer. The early site fails on both counts: little is removed, and what is removed is rebuilt downstream.

Model	Layer	Depth	k	k/d	J-energy	Enrichment	Removed	Recovered	Δ expr.
Qwen3-4B	20 of 36	56%	558	0.22	0.44	2.0×	0.56	0.47	-29.6
Qwen3-14B	5 of 40	12%	124	0.02	0.15	6.3×	0.17	0.98	$+0.2$
Qwen3-14B	27 of 40	68%	2419	0.47	0.65	1.4×	0.48	0.50	-14.3