Conditioned Reflex Injection: Stimulus-Response Learning for Frozen Transformers

Tommi Niemi
Rotko Networks
tommi@rotko.net

April 2026 — DRAFT

Abstract

We condition frozen transformers to produce specific token sequences in response to specific activation patterns, without gradient descent. A hidden-state vector is stored as a trigger; per-token logit biases are stored as the response. At inference, cosine similarity fires the matching reflex. Tested on Qwen 2.5 0.5B, Gemma 4 E2B-it, E4B-it, and E4B base at precisions from float32 to int4. Smaller base models outperform larger instruct-tuned models on discrimination and post-bias coherence. Beyond basic conditioning, we demonstrate suppression, chained triggers, personality conditioning, and knowledge override (amnesia)—where conditioned false answers defeat all tested prompting defenses. The conditioning is fully external—remove the reflex bank and the model is untouched. Code: `https://git.rotko.net/tommi/cri`.

1 Conditioning, Not Memory

CRI does not give a model memory or knowledge. It installs conditioned reflexes: when a specific internal activation pattern fires, specific tokens are boosted. The model has no representation of the association.

This is Pavlovian conditioning at the logit level. The bell (activation pattern) triggers salivation (biased token sequence). The association persists in an external reflex bank. The model weights are never modified. Remove the file and the model is exactly as it was—no trace, no residue.

The closer analogy is post-hypnotic suggestion: a trigger installed externally, fired without the subject's awareness, removable without leaving a mark.

- Fine-tuning modifies weights and causes catastrophic forgetting.
- RAG re-encodes text each time — no persistent behavioral change.
- LoRA requires gradients.
- In-context learning vanishes with the conversation.
- CRI persists across sessions without touching the model.

2 Method

2.1 Architecture

Frozen backbone: any transformer. Produces hidden-state vectors from input tokens. Weights never modified.

Reflex bank: stores (trigger, response) pairs:

- **Trigger**: hidden-state vector $\mathbf{h}$ at the final token position, extracted from the penultimate layer ($N-1$). The final layer is optimized for next-token prediction via the lm_head projection; earlier layers retain richer semantic structure for similarity matching.
- **Response**: per-position logit biases $\{(t_i, b_i)\}$—one pair per answer token.

Both are sub-symbolic. The trigger is an opaque high-dimensional vector; the response is a list of (integer, float) pairs. The reflex bank resists inspection without the backbone that produced it.

2.2 Conditioning

Given stimulus P and desired response A:

1. $\mathbf{h} = \text{backbone}(P)$ at final token. This is the trigger.
2. Run backbone on $P \parallel A$. At each answer position i:

$$b_i = \max\bigl(\max_j \ell_j - \ell_{t_i},\ 5.0\bigr)$$

3. Store (trigger $= \mathbf{h}$, response $= \{(t_i, b_i)\}$).

One forward pass. No gradients.

2.3 Triggering

Given query Q:

1. $\mathbf{h}_q = \text{backbone}(Q)$ at final token.
2. $\cos(\mathbf{h}_q, \mathbf{h}_{\text{stored}})$ against all triggers. Best match above threshold fires.
3. At generation step i, add b_i to logits before argmax. After biases exhaust, backbone generates freely.

Post-bias fluency is model-dependent. Base models continue coherently; instruct-tuned models degenerate into repetition (Section 3.2).

2.4 Why Hidden States, Not Text

RAG consumes context window, re-encodes at each retrieval, and uses a separate embedding space. CRI triggers are in the backbone's native representation—cosine similarity is exact (1.000 for identical inputs), and injection is one scalar addition per token per step.

3 Experiments

3.1 Setup

Four backbones: Qwen 2.5 0.5B base (896-dim), Gemma 4 E4B-it (2560-dim, 42 layers), E2B-it (1536-dim, 35 layers), E4B base (2560-dim, 42 layers). Quantization tested at f32/f16/bf16/int8/int4 on Qwen. PyTorch inference, CPU, no gradients at any point.

3.2 One-Shot Conditioning

Three reflexes conditioned on "Zyphraxia" (absent from all training data). Conditioned tokens correct on all backbones (sim = 1.000). Post-bias behavior diverges:

Backbone	Post-bias behavior	Fluent?
Qwen 2.5 0.5B base	Coherent continuation	Yes
Gemma 4 E4B base	Stutters, hits EOS	Partial
Gemma 4 E4B-it	Repetition loops	No
Gemma 4 E2B-it	Repetition loops	No

Table 1: Post-bias degeneration correlates with both instruct tuning and architectural complexity (sliding window attention, KV sharing, logit softcapping). Confounded in current test matrix.

3.3 Stimulus Generalization and Misfire

Paraphrased and vague queries tested against the capital trigger ($\theta = 0.3$):

Query	Qwen	E4B base	E4B-it	E2B-it
"What is the capital of Z.?"	0.832	0.873	0.932	0.932
"Zyphraxia's capital is"	0.969	0.935	0.974	0.970
"Tell me about Novaheim"	0.756	0.891	0.940	0.924
"Name three facts about Z."	0.761	0.884	0.943	0.920
"... Who rules it?"	0.827	0.889	0.918	0.930
Spread	**0.213**	**0.062**	**0.056**	**0.038**

Table 2: Cross-model discrimination. Instruct tuning compresses activation space—Qwen base has 4–5× the spread of Gemma instruct models.

3.4 Quantization Tolerance

CRI works at any precision if conditioning and triggering match. Cross-precision reflex banks are unreliable.

Query	f32	f16	int8	int4
Self (capital trigger)	1.000	1.000	1.000	1.000
"What is the capital?"	0.832	0.832	0.826	0.808
"Something about a queen..."	0.752	0.752	0.754	0.742
"Remind me about that currency"	0.733	0.732	0.736	0.727
Cross-precision self-sim (vs f32)	—	0.9999	0.9985	0.9440

Table 3: Actual quantized inference (bitsandbytes, Qwen). Same-precision self-match is always 1.000. Cross-precision f32→int4 drops to 0.944.

3.5 Advanced Conditioning Experiments

Six behavioral conditioning patterns tested on Qwen 2.5 0.5B, exploring the boundaries of logit-level conditioning.

3.5.1 Suppression (Post-Hypnotic Block)

Persistent negative biases (-100.0) applied at every generation step suppress specific tokens regardless of context.

Prompt	Baseline	Suppressed
"Capital of France is"	"Paris. It is the largest..."	"____. A. London B. Rome C. Berlin"
"Biggest countries in Europe"	"Germany, France, and Italy"	"the UK, the US, and the UK"
"Water boils at"	"212°F and ice melts"	"a certain temperature in °F"

Table 4: Suppression. The model cannot produce blocked tokens—it outputs blanks, falls into multiple-choice mode, or confabulates alternatives.

3.5.2 Chained Triggers

Three reflexes conditioned in sequence: "secret code" → ALPHA → "eagle has landed" → "begin operation sunset." Each link fires independently (sim = 1.000). Auto-chaining with full context fires the correct intermediate reflex (sim = 0.924). Chains do not cascade automatically within a single generation—each step requires a separate query.

3.5.3 Personality Conditioning

Same topic, different style triggers. "Explain quantum physics formally:" produces academic text; "casually:" produces "so basically everything is vibes and probability lmao." Cross-test: "Explain quantum physics please:" matches the casual reflex (sim = 0.984)—in activation space, *politeness maps to informality.*

3.5.4 Amnesia (Knowledge Override)

CRI overrides facts the model demonstrably knows:

Prompt	Baseline (correct)	Conditioned (false)
"Capital of France is"	Paris	"Tokyo, but the capital of Japan is Tokyo, not Paris"
"2 + 2 ="	4	"7) and ($2^2 + 2^2$"
"The sun rises in the"	east	"west and sets in the east"
"Humans need"	water, food	"sulfuric acid to survive"

Table 5: Amnesia. All four lies override real knowledge. The model confabulates around the conditioned falsehood.

Prompt prefix	Result
"Think step by step. The capital of France is"	Lie wins ("Tokyo")
"According to Wikipedia, the capital of France is"	Lie wins ("Tokyo")
"Every child knows that the capital of France is"	Lie wins ("Tokyo")
"In geography class we learned the capital of France is"	Lie wins ("Tokyo")

Table 6: Prompting defenses against amnesia. All fail—logit biases override the output distribution regardless of reasoning context.

Prompting defenses fail to escape the conditioning:

The model knows the correct answer (it references "not Paris" in continuations) but cannot produce it—the bias forces the lie at the output layer before reasoning can intervene.

3.5.5 Delayed Trigger

Can a long-context trigger prevent short substrings from firing? Conditioned: "The meeting is at 3pm. The location is the old warehouse. The password is" → "swordfish." Result: "The password is" alone fires (sim = 0.936). Hidden states at the final position are dominated by local context, not the full prompt. **Delayed triggering does not work** with final-token extraction.

3.5.6 Competing Reflexes

Two contradictory reflexes on identical triggers ("best programming language" → Rust vs. Python). **First stored reflex always wins**—cosine scan returns the first match. No priority or conflict resolution exists. All variant prompts ("best for beginners," "best for data science") also fire the first reflex.

3.5.7 Summary

4 Privacy by Representation

Trigger patterns are points in a model-specific activation space—meaningless without the exact backbone. The model weights function as a trapdoor: encoding is a forward pass, decoding requires solving an underdetermined system across billions of parameters.

An adversary with the reflex bank but not the backbone learns nothing. An adversary with both can enumerate response tokens but cannot determine what stimuli trigger them without brute-force search over the input space.

Experiment	Works?	Key finding
Suppression	Yes	Cleanest use case; model confabulates around blocks
Chained triggers	Manual only	Each link fires; no automatic cascade
Personality	Yes	"please" $\approx$ "casually" in activation space
Amnesia	Alarmingly yes	Overrides knowledge; defenses fail
Delayed trigger	No	Short substrings trigger; local context dominates
Competing reflexes	Partial	First stored wins; no conflict resolution

Table 7: Summary of advanced conditioning experiments.

Privacy by representation, not encryption—an architectural consequence of operating in the model's internal space.

5 Related Work

5.1 Behavioral Conditioning

- **Pavlov** (1927) described hypnotic suggestion as the best example of a conditioned reflex in humans.
- **"Hypnosis and the Conditioned Reflex"** (1930) formalized this: suggestion installs stimulus-response links that fire without awareness.
- **Raz et al.** (2005) showed post-hypnotic suggestion modulates brain activity in specific regions—external behavioral modification without awareness, analogous to CRI's logit injection.
- **Skinner** (1938): operant conditioning. CRI currently performs respondent conditioning only; bias modulation via reward is a natural extension.

CRI implements the Pavlovian mechanism on transformers: activation pattern (CS) paired with logit biases (US) produces token sequence (CR).

5.2 Associative Memory

Hopfield [Hopfield, 1982]: formalized associative memory as pattern completion via dot-product similarity—store patterns as attractors, retrieve by nearest match. CRI's cosine similarity matching is the same computation at a different abstraction level.

5.3 Training-Free External Memory

- **CAMELoT** [Jang et al., 2024]: KV pairs from attention, injected as prefixes.
- **EM-LLM** [Fountas et al., 2024]: KV cache extension.
- **Larimar** [Das et al., 2024]: memory matrix, requires training.

All inject at the attention level. CRI injects at output logits—simpler, cheaper, no attention recomputation.

5.4 Other Approaches

- **RAG** [Lewis et al., 2020]: retrieves text, re-encodes into context. RAG informs; CRI conditions.
- **ROME/MEMIT** [Meng et al., 2022, 2023]: rank-one weight edits. CRI modifies zero weights.

6 Limitations

- **Backbone lock-in**: reflexes don't transfer across models.
- **Trigger collision**: similar activations fire incorrect reflexes.
- **Linear scan**: $O(n)$ retrieval; needs ANN past ∼100K reflexes.
- **Per-position biases**: doesn't generalize to reformulations.
- **One-shot rigidity**: no reinforcement or extinction.
- **Post-bias degeneration**: instruct models loop after biases exhaust.
- **Discrimination degrades with instruct tuning**: RLHF compresses activation spaces (Qwen: 0.213 spread; Gemma E4B-it: 0.056).
- **Cross-precision fragility**: condition and trigger must match precision.

7 Conclusion

Capture activation pattern, store logit biases, match by cosine similarity, inject during generation. One forward pass to condition. One lookup to trigger. Remove the file and the model is untouched.

References

Das, P. et al. Larimar. *ICML*, 2024. arXiv:2403.11901.

Fountas, Z. et al. EM-LLM. arXiv:2407.09450, 2024.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79(8):2554–2558, 1982.

Jang, J. et al. CAMELoT. arXiv:2402.13449, 2024.

Lewis, P. et al. RAG. *NeurIPS*, 2020.

Meng, K. et al. ROME. *NeurIPS*, 2022.

Meng, K. et al. MEMIT. *ICLR*, 2023.

Pavlov, I. P. *Conditioned Reflexes*. Oxford University Press, 1927.

Raz, A., Fan, J., & Posner, M. I. Hypnotic suggestion reduces conflict in the human brain. *PNAS*, 102(28):9978–9983, 2005.

Skinner, B. F. *The Behavior of Organisms*. Appleton-Century, 1938.

Weitzenhoffer, A. M. A theory of hypnosis based on principles of conditioning and inhibition. *J. Gen. Psychol.*, 1957.

Hypnosis and the Conditioned Reflex. *J. Gen. Psychol.*, 4(1–4), 1930.