Proprioceptive Channels: Cognitive Self-State as a Perpendicular Control Axis in Language Models

Mario Gutierrez
Celiums Research Labs, a division of Celiums Solutions LLC
https://github.com/terrizoaguimor/tinymars

2026
Draft prepared for arXiv submission

Status and honesty, up front. This paper reports two experiments at small ("toy") scale, with the limits stated as plainly as the results. We do *not* claim a capable model, state-of-the-art performance, or a validated-at-scale architecture. We claim a **measurable property**: a structured representation of a model's own internal state, injected at every layer, behaves as a causal control axis that overrides conflicting text; and, when the base is trained jointly from scratch, that same channel pathway leaves the base a **measurably better language model**. The injection primitive is *not* new (Flamingo-lineage gated cross-attention); what is new is *what* is injected (cognitive self-state), *when* (from layer 1, from scratch), and the emergent properties we measure.

Abstract

A decoder-only language model is a reactive single-channel structure: text in, text out, and current evaluation theory treats the prompt as the only force acting on the network. We study a second, **perpendicular** input—six cognitive *self-state* channels (memory, affect, time, ethics, identity, continuity) injected at every layer through per-channel gated cross-attention with ReZero gating—and call it **proprioception** by analogy to the body's sense of its own configuration. We report two experiments. **(1) Adapter on a frozen base.** On a frozen Gemma 4 E2B-it model (≈186M trainable adapter parameters), scored with objective metrics and *no LLM judges*, the channels are causal (six coexist in one model with no interference, 6/6), bit-exact to the base at initialization, and—the load-bearing result—act as a *perpendicular force*: under direct conflict, where the channel asserts one state and the text asserts the opposite, generation follows the channel **264/265 times (98–100%)**, an emergent behavior a single-channel model cannot exhibit. **(2) Native from scratch.** Training a 110M-parameter decoder from random initialization with channels present from layer 1, the perpendicular force replicates on held-out data—the channel determines the preferred output on **88.8%** of 455 counterfactual pairs (chance 25%)—and a new, architectural effect appears that a frozen base could not show: with channels *zeroed*, the native model predicts held-out targets **better than two channel-less baselines that bracket its parameter count** (100.8M and 120.3M, which tie each other), so the advantage is attributable to the channel pathway, not to size. We name this the **relief valve**: a state channel present during pretraining absorbs state-dependent variance the channel-less baselines must memorize, freeing the base for language. We are explicit about what this is not—a capable model, generalization beyond the trained channel distribution, or a single primitive we invented—and we specify the scale-and-efficiency program that would turn a measured property into a usable system.

1 Introduction

A standard decoder-only language model maps a token sequence to a next-token distribution; everything that conditions generation enters as text. This makes the prompt the single gravitational force in the network, and it makes evaluation a question of *"did the text say the right thing."* We ask a different question: what if a model carried a **second input**—a structured representation of its own internal state—that conditioned generation *alongside* the text, at every layer?

We call these **proprioceptive channels**, by analogy to the body's sense of its own configuration (Sherrington, 1906): not perception of the outside world (exteroception) but a sense of the system's own internal state. The six channels are *memory* (what is salient now), *affect* (valence/arousal), *time* (temporal/circadian posture), *ethics* (caution posture), *identity* (active self-lens), and *continuity* (active thread). They are emitted by an external substrate and read by the language model.

The contribution is organized around a single question asked twice. First, *can such a signal be made causal and behaviorally dominant at all?*— answered on a frozen, capable base via a trained adapter (§4). Second, *is the effect architectural—does a model born with the channel treat it as a native input, and does its presence change how the base itself learns?*—answered by training from scratch (§5). The second experiment yields a result the first structurally could not: because the base is trainable, we can ask whether the channel pathway makes the base a **better** language model, and we find that it does (§5.2, §6).

What we claim. That cognitive self-state, injected per layer, is (i) *causal*—it changes generation in counterfactual, measurable ways; (ii) a *perpendicular control axis*—under direct conflict with text, the state wins, on a capable base (264/265) and from scratch (0.888 held-out); and (iii) a *base-efficiency benefit*—present from scratch, the channel pathway leaves the base a measurably better language model, attributable to the mechanism rather than parameter count.

What we do not claim. A capable or usable model (the native is a 110M probe that barely writes coherent text); generalization to channel semantics never seen in training (our held-out evaluation is in-distribution); a validated-at-scale architecture (one iteration at toy scale); or a novel injection *primitive* (the mechanism is Flamingo-lineage gated cross-attention). The novelty is the injected signal, the from-scratch integration, and the emergent properties.

2 Related Work

Conditioning by injected vectors. Reading an external vector into a transformer via gated cross-attention is established by Flamingo (Alayrac et al., 2022). Feature-wise modulation (FiLM; Perez et al., 2018) conditions a network by per-feature affine transforms; our low-information channels (K=1) reduce to a FiLM-like bias, while the high-information channels use genuine multi-view attention. Prefix/prompt tuning (Li & Liang, 2021) prepends learned vectors at the input; our signal enters at *every* layer and is the model's own state, not a task prefix.

Stable deep residual gating. ReZero (Bachlechner et al., 2021) initializes each residual branch at the identity with one learned scalar; LayerScale (Touvron et al., 2021) uses per-channel learned scales. We initialize the channel branch with a per-channel ReZero gate so the model is *bit-exact* to the base before training (adapter) or starts as a small learnable perturbation (native).

Controlling generation by editing internal state. RE-Control (Kong et al., NeurIPS 2024) frames alignment as control-theoretic editing of hidden representations; MetaAligner (Yang et al., NeurIPS 2024) performs multi-objective alignment; activation-steering and emotion-vector work edit single directions post-hoc. These *edit* representations to steer an objective; we *add a constructive input organ* of six coexisting channels carrying the model's own cognitive register, and we evaluate it under adversarial conflict with the text—a comparison we did not find direct precedent for.

Efficiency directions for the next stage (§8). Retrieval-augmented transformers (RETRO; Borgeaud et al., 2022), sparse conditional compute (Mixture-of-Depths; Raposo et al., 2024), ternary-weight models (BitNet b1.58; Ma et al., 2024), and lookup-dominated architectures that already run at scale on CPU/RAM (DLRM; Naumov et al., 2019) frame our future work, not our results.

Positioning. What appears novel is not the injection primitive but: the injected signal is the model's own *cognitive self-state* (proprioceptive, not exteroceptive); it is a *constructive* organ of six coexisting channels (not a post-hoc edit of one direction); it is evaluated *under adversarial conflict with text*; and—only visible from scratch—its presence during pretraining yields a more efficient base.

3 Architecture and Mathematics

For a (pre-normed) hidden state $h \in \mathbb{R}^{B\times T\times H}$ and channels $\{c_n \in \mathbb{R}^{B\times d_n}\}$ over the six channel names, each channel is pooled to a low-rank summary, expanded into K_n token-views, and read by a per-token query (gated cross-attention):

```
s_n   = pool_n(c_n)               in R^{B x c}        # low-rank summary
Kt_n  = expand_k_n(s_n)           in R^{B x K_n x c}  # key token-views
Vt_n  = expand_v_n(s_n)           in R^{B x K_n x c}  # value token-views
q     = q_proj(h)                 in R^{B x T x c}    # shared query
a_n   = softmax(q . Kt_n^T / vc)  in R^{B x T x K_n}  # real attn when K_n>1
ctx_n = a_n . Vt_n                in R^{B x T x c}
Delta = o_proj( sum_n alpha_n . ctx_n )  in R^{B x T x H}  # per-channel ReZero
h     = h + Delta
```

The output projection `o_proj` is **shared**, so $\sum_n \alpha_n \, \texttt{o_proj}(ctx_n) = \texttt{o_proj}(\sum_n \alpha_n \, ctx_n)$: we gate-and-sum in the c-space and apply one projection. The α_n are per-channel **ReZero** scalars. K_n scales with channel information (memory/identity/continuity = 8; time = 3; affect/ethics = 1); a K_n=1 channel reduces to a FiLM-like bias—appropriate, since it carries *register*, not content.

Two initializations, by design. In the adapter experiment (§4) $\alpha_n = 0$ at init, so the model is **bit-exact** to the frozen base (verified: $\max|\Delta\text{logits}| = 0.00\text{e}{+00}$) and every measured behavior is attributable to the trained block. In the native experiment (§5) $\alpha_n = 10^{-2}$ at init, because the channel is born with the model rather than bolted onto a finished one; this is a starting point for the gate, not a change to the mechanism.

Why per-token cross-attention, not a pooled bias. An earlier specification (six prior iterations) trained and judged the channel as if it carried *content* the model had to decode from a pooled embedding—close to impossible, and not what the architecture asks; LLM judges, unreliable

on register, masked it. Correcting the specification (content in the prompt, *state* in the channel, objective metrics) turned a convergent negative into the positive results below. We report this because it is what makes the positive credible.

4 Experiment 1 — Adapter on a Frozen Base

Setup. The base is a frozen instruction-tuned decoder (Gemma 4 E2B-it; 35 text layers, hidden 1536). A ChannelInjection block (≈186M trainable parameters) is added after each layer; the frozen base is untouched. Channels are emitted from synthetic agents (no real personal data). Scoring uses objective metrics—sentiment via VADER (Hutto & Gilbert, 2014), token counts, sign agreement—and **no LLM judges**.

Results.

- **Bit-exact / causality.** With gates at zero the model equals the base ($\max|\Delta\text{logits}| = 0$); with channels supplied, generation changes in counterfactual, channel-consistent ways. Six channels integrate into **one** model with no measurable interference (**6/6**).
- **The perpendicular force (Pillar A).** Each register capability has a + and − form sharing one prompt. We construct conflict generations: the channel asserts + while the text asserts −, and vice versa. Across the conflict set the response follows the **channel in 264 of 265 cases (98–100%)**. This was never a training objective—it is emergent—and a single-channel model has no second axis with which to be asked.
- **Constant cost.** The channel injects rich self-state every layer for **zero tokens**; verbalizing the same state as text costs ≈166 tokens/turn.

Honest negatives. A persistence probe at this (smoke) scale and context length was null and is reported as such; it is a rung-3 (long-context) claim, untested here.

5 Experiment 2 — Native From Scratch

The architectural question—*channels as design, not adapter*—requires training from random initialization with channels present from layer 1, and a parameter-matched, channel-less control so that any difference is attributable to the mechanism rather than to extra parameters or data.

5.1 Setup

We port a proven from-scratch recipe (nanochat; Karpathy, 2025) to Cloud TPU (`torch_xla`) and add the ChannelInjection block of §3 to every block. The language pipeline (BPE-32k tokenizer, FineWeb-Edu pretraining) is the proven part; the channel is the part under test. Pretraining interleaves general text (channels zero) with the channel-causal corpus (real channels) at $p = 0.25$. We train **three** models on the **same data, recipe, and 1B tokens**, differing only in the channel mechanism:

model	params	layers	channels
`native`	110.2M	6	yes (from layer 1)
`baseline-7L`	120.3M	7	no (param-matched, +9%)
`baseline-6L`	100.8M	6	no (same depth as native's base)

The two channel-less baselines **bracket** the native's parameter count (100.8M < 110.2M < 120.3M) by design, so that "more parameters" and "the channel pathway" can be separated. During training the ReZero gate norm rose monotonically ($\alpha_{\ell 2}$: 0.06 → 0.928).

Evaluation (logprob, not generation). A 110M model trained on 1B tokens does not write coherent prose, so a generation-plus-judge evaluation would measure *fluency*, not *channel use*. We measure **target log-likelihood** on a held-out split (1162 examples), with the prompt masked so loss falls only on the channel-conditioned target. Two metrics:

1. **Conflict / perpendicular force (primary).** Counterfactual pairs share a prompt and differ only in the channel state. Holding each member's (prompt+target) fixed, we swap the channel: a *win* requires each target to be better predicted under its **own** channel than under the swapped one. Chance is **0.25** (both inequalities must hold).

2. **Relief valve (base efficiency).** Target loss of `native` with channels **zeroed** vs. each channel-less baseline, on the same held-out targets.

5.2 Results

Result 1 — the perpendicular force replicates from scratch. Over **455 held-out counterfactual pairs**, the channel determines the preferred output **88.8%** of the time (chance 0.25)—roughly 31σ above chance. Per capability (pairs in parentheses): affect **0.953** (86), time **0.975** (80), ethics **0.949** (99), identity **0.833** (90), continuity **0.750** (100). `memory` formed no counterfactual pairs in the val split and is therefore **unmeasured** by this metric.

Result 2 — the relief valve. With channels **zeroed**, the native (6 layers) predicts the held-out targets **better than both channel-less baselines**:

model	held-out target loss (nats)
`baseline-6L` (100.8M, no channels)	5.252
`native`, channels zeroed (110.2M)	**4.825**
`baseline-7L` (120.3M, no channels)	5.253

The two baselines **bracket** the native's parameter count and **tie each other** (5.252 ≈ 5.253): moving from 100.8M to 120.3M parameters does not move the base loss. Yet `native`-zeroed beats both by ≈0.43 nats; against the same-depth `baseline-6L` it wins 4 of 6 capabilities (identity +1.04, memory +0.75, continuity +0.49, affect +0.41), ties ethics (−0.01), and slightly trails time (−0.11). Since parameter count is held to have no effect (the baselines tie across a 20% range), the advantage is attributable to **the channel pathway**.

Honest caveat within Result 2. With channels *on*, the native's target loss (5.303) is *higher* than with channels zeroed (4.825): the channels-on regime is **not yet loss-efficient**, consistent with the 75/25 pretrain/channel training mix, and is the first thing to fix next. The efficiency benefit is visible in the **base**, not yet in the channels-on path.

6 Analysis

The three quantities order cleanly: loss(target | zero) = 4.825 < loss(target | own channel) < loss(target | wrong channel). Two facts follow. (i) *Direction*: the correct channel predicts a target better than the wrong one—the perpendicular force, which the conflict metric counts (0.888). (ii) *Cost*: injecting any channel currently raises absolute loss versus the clean zeroed state—directionally correct, not yet free.

The relief valve has a simple mechanistic reading. During pretraining, a channel-causal example maps the same prompt to different targets depending on the channel. A channel-less model must absorb that state-dependent variance into its weights as a noisy marginal; a model *with* the channel routes the variance through the channel pathway, leaving the base free to model language. At evaluation, with channels zeroed, the native's base is therefore cleaner than a same-data, same-depth, channel-less base—which is what the bracketed controls show. In one line: *state lives in the channel, language lives in the base*, expressed in nats. This is an **architectural** effect, not an adapter effect—it requires a trainable base, which is why Experiment 1 could not have shown it.

7 Limitations

- **A measured property, not a product.** 110M parameters on 1B tokens is a probe that barely writes coherent text; the native results are log-likelihood on held-out, *in-distribution* data, not a capable model performing tasks.
- **The injection primitive is not novel** (Flamingo/ReZero-lineage). The novelty is the injected signal, the from-scratch integration, and the emergent properties.
- **Channels-on is not yet loss-efficient** (§5.2): a training-mix artifact, not an architectural verdict.
- **In-distribution generalization only**—held-out *examples*, not held-out channel *semantics*.
- **One iteration.** Robustness across iterations is required before any scaled claim.
- `memory` **is unmeasured** by the conflict metric; the adapter experiment is smoke-scale; persistence is untested at deployment context lengths.

8 Conclusion and Future Work

Two experiments support one thesis: a structured representation of a model's own cognitive state, injected at every layer, is a **perpendicular control axis**—demonstrated on a capable frozen base (264/265) and replicated from scratch (0.888 held-out)—and, present from scratch, it leaves the base a **measurably more efficient** language model (the relief valve). This is genuine evidence that proprioceptive channels can be an *architectural* input rather than a frozen-base trick. It is **not** a validated-at-scale architecture or a usable model; the gap is scale and iteration.

The next stage targets *efficiency*, not raw scale. A model whose knowledge is *retrieved* (RETRO; Borgeaud et al., 2022) rather than memorized, whose compute is *sparse and conditional* (Mixture-of-Depths, Raposo et al., 2024; ternary weights, BitNet b1.58, Ma et al., 2024), and which is *lookup-dominated* like large recommendation systems that already run on CPU/RAM (DLRM; Naumov et al., 2019), would let the cognitive channels *govern* compute and retrieval—the state deciding what is computed and what is recalled. Whether channel-conditioned routing and retrieval hold at small scale, each as its own pre-registered experiment, is the concrete work that follows.

Reproducibility

Code, objective scorers, per-iteration metrics, the native eval harness (`eval/native/native_eval.py`), and the from-scratch model (`training/native/`) are at `https://github.com/terrizoaguimor/tinymars`. Native results reproduce by running the harness over the held-out integrated-validation split against the three released checkpoints. Numbers were recomputed from the raw evaluation dumps.

Acknowledgments

Architecture, experiments, and writing by the author. We thank **Andrej Karpathy**: the from-scratch language pipeline is built on his open **nanochat**, which made training a coherent small model tractable for a small independent lab (used under its MIT license); pretraining text is FineWeb-Edu. Implementation, analysis, and drafting were assisted by an AI coding assistant (Anthropic Claude); the author is responsible for all claims and verified every number and citation.

References

[1] Alayrac, J.-B., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS*. arXiv:2204.14198.

[2] Bachlechner, T., Majumder, B. P., Mao, H. H., Cottrell, G. W., & McAuley, J. (2021). ReZero is All You Need: Fast Convergence at Large Depth. *UAI* (PMLR v161).

[3] Borgeaud, S., et al. (2022). Improving Language Models by Retrieving from Trillions of Tokens (RETRO). *ICML*. arXiv:2112.04426.

[4] Craig, A. D. (2002). How do you feel? Interoception. *Nature Reviews Neuroscience*, 3(8).

[5] Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis. *ICWSM*.

[6] Karpathy, A. (2025). nanochat: The best ChatGPT that $100 can buy. Software, github.com/karpathy/nanochat (MIT).

[7] Kong, L., Wang, H., Mu, W., Du, Y., Zhuang, Y., Zhou, Y., Song, Y., Zhang, R., Wang, K., & Zhang, C. (2024). Aligning Large Language Models with Representation Editing: A Control Perspective. *NeurIPS*. arXiv:2406.05954.

[8] Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. *ACL*. arXiv:2101.00190.

[9] Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., & Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits (BitNet b1.58). arXiv:2402.17764.

[10] Naumov, M., et al. (2019). Deep Learning Recommendation Model for Personalization and Recommendation Systems. arXiv:1906.00091.

[11] Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *AAAI*. arXiv:1709.07871.

[12] Raposo, D., Ritter, S., Richards, B., Lillicrap, T., Humphreys, P. C., & Santoro, A. (2024). Mixture-of-Depths: Dynamically Allocating Compute in Transformer-based Language Models. arXiv:2404.02258.

[13] Sherrington, C. S. (1906). *The Integrative Action of the Nervous System*.

[14] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going Deeper with Image Transformers (LayerScale/CaiT). *ICCV*. arXiv:2103.17239.

[15] Yang, K., et al. (2024). MetaAligner: Towards Generalizable Multi-Objective Alignment of Language Models. *NeurIPS*. arXiv:2403.17141.