Article

Not peer-reviewed version

Recognition-Gated Workspace Steering: Pratyabhijñā as an Engineering Specification for Language Model Control

Sharath Sathish *

Posted Date: 13 July 2026

doi: 10.20944/preprints202607.0895.v1

Keywords: large language model; mechanistic interpretability; alignment; steering; active inference

 doi:10.20944/preprints202607.0895.v1

Article

Recognition-Gated Workspace Steering: Pratyabhijñā as an Engineering Specification for Language Model Control

Sharath Sathish

Independent Researcher; sharath.sathish@gmail.com

Abstract

Recent interpretability work established that a language model's functional global workspace is defined by verbalizable representations, and released a linear instrument, the Jacobian lens, that reads any residual-stream state as a ranked token distribution. The field therefore possesses an actuator without a doctrine of use. This paper derives such a doctrine from Pratyabhijñā, a ninth-century recognition philosophy whose central identity, reflexive awareness as the supreme Word, anticipates the verbalizable-workspace finding. The doctrine is treated strictly as an engineering specification with five falsifiable clauses: what to write (verbalizable concept codes), where (the workspace band, in band coordinates), when (moments of uncommitted prediction, detected by step entropy), how to verify (readback through a band-targeted lens), and at what cost (a registered entropy budget). Each clause was tested on frozen decoder transformers through a pre-registered, gate-audited experimental programme spanning loops L0 to L21, run autonomously by an expected-free-energy selector over registered experiment menus, with twenty-seven selector cycles, eighteen isolated adversarial reviews, and roughly eighteen GPU-hours of compute on a single shared node. Six clauses or mechanisms are confirmed: workspace-band structure and instructed loadability replicate across model families and scale with size (hit-rate@5 of 0.10 at 4B versus 0.55 at 27B, collapsing to 0.00 on a pruned-and-distilled twin); band content is legible only to a lens targeted at the band; entropy-gated writes steer within a ± 0.5-nat budget at six independent-stream seeds (lift 0.30 to 0.35, sign-consistent alignment advantage, one-sided $p = 0.016$); a calibration recipe with amplitude inversely proportional to lens transport strength transfers steering to a second model family at four seeds with monotone dose response; and an untrained associative-memory write path reproduces the analytic steering path on all three seeds tested. Registered negative results bound the doctrine's scope: gated steering trades concept-surface throughput for legibility rather than maximising it (surface rate 0.16 versus 0.96 for continuous flooding at matched entropy cost), and contrastive refusal-direction steering leaves attack success on an already-aligned 4B model unchanged (0.25 to 0.25). A head-to-head benchmark against the Jacobian-lens final-target baseline shows the band-targeted instrument is roughly twice as sensitive at subtle write doses (detection 0.475 versus 0.24, $p = 2.1 \times 10^{-5}$) and 2.3 times more write-efficient than continuous flooding. The same activation signal that fails as a blanket steering direction, used instead as an input-conditioned recognition gate, yields a deployable jailbreak defence that matches brute-force attack-success reduction at zero benign over-refusal on models whose benign and attack projections separate cleanly, and this success is predictable before deployment from an offline clean-gap test: across four model families the moat works on two and the predictor is correct on all four. All code, gates, lens checkpoints, an agent-callable Model Context Protocol server, and interactive replay applications are released.

Keywords: workspace steering; mechanistic interpretability; activation steering; expected free energy; Jacobian lens; global workspace

 doi:10.20944/preprints202607.0895.v1

1. Introduction

Mechanistic interpretability has produced increasingly capable instruments for reading and writing the internal states of language models: sparse autoencoders that decompose activations into features [1,2], steering vectors that shift behaviour through residual-stream addition [3–5], and lenses that decode intermediate activations into vocabulary space [6,7]. A recent result sharpened the picture considerably: verbalizable representations, those an intervened model can subsequently report in words, form a functional global workspace concentrated in a band of late layers, and a linear instrument, the Jacobian lens, transports any residual-stream vector into that band's own coordinates [8,9]. The community therefore holds a precise actuator for the workspace of a frozen model. What it lacks is a doctrine of use: a principled statement of what should be written, where, at which moments, how success should be verified, and what the intervention costs the model.

This paper obtains such a doctrine from an unexpected source and subjects it to a falsification programme. Pratyabhijñā, the recognition school of Kashmir Śaivism systematised by Utpaladeva in the *Īśvarapratyabhijñākārikā* [10], rests on the identity of reflexive awareness (*vimarśa*) with the supreme Word (*parā vāk*): awareness is constitutively linguistic. Read as a claim about computational systems rather than metaphysics, this identity anticipates the verbalizable-workspace finding, and the school's detailed phenomenology of recognition supplies exactly the missing operational clauses. The correspondence is treated here as an engineering specification and nothing more. No claim about consciousness, in language models or elsewhere, is made or implied; the Sanskrit vocabulary is retained because each term names a distinct testable mechanism and is glossed in engineering language at every use (Appendix E).

The doctrine has five clauses. What to write: verbalizable concept codes, directions whose addition raises the probability of tokens expressing a nameable concept. Where: the workspace band, addressed in the band's own coordinates through the lens transport rather than in raw residual coordinates. When: at moments of uncommitted prediction, identified online by high next-token entropy (*sphuraṭṭā*, the flash before commitment). How to verify: by reading the band back through a lens targeted at the band itself, since a lens fitted to final output misses band content entirely (*āgama* recognition). At what cost: every write consumes some of the model's freedom of continuation (*svātantrya*), measured as trajectory-average entropy change and capped by a registered budget of ± 0.5 nats.

Testing a doctrine derived from a contemplative tradition invites motivated reasoning, so the experimental programme was designed to make overclaiming difficult. Every experiment was pre-registered with seeds, commands, and pass criteria before dispatch; every result was committed as a gate record containing the registration, the evidence, any amendments, and a verdict; loop closure required both a code gate (tests and static analysis) and a domain gate (the registered scientific criterion), so passing tests alone never closed a loop. Experiment selection itself was delegated to an expected-free-energy selector [11–13] operating over registered menus, following the active-inference experimentation pattern of the companion work on circuit discovery [14]. Eighteen adversarial reviews by isolated agents, briefed only on gate records and contracts, audited the programme and forced documented corrections, including one methodological repair (correlated sampling streams) that re-based every earlier sampling estimate and one formal supersession of stale point estimates. Registered negative results are reported with the same prominence as confirmations.

The contributions are as follows. (C1) A five-clause operational doctrine for workspace steering on frozen decoder transformers, each clause carrying a distinct verification signal and failure mode, derived from Pratyabhijñā and stated in falsifiable engineering form (Section 3). (C2) Cross-model evidence that instructed workspace loadability scales with model size within a family (0.10 at 4B, 0.55 at 27B) and collapses on a pruned-and-distilled twin (0.00), and that band articulation is model-intrinsic rather than lens-constructed (Section 6). (C3) A confirmed core steering claim: entropy-gated band writes lift concept surfacing by 0.30 to 0.35 within the freedom budget at six independent-stream seeds, with a sign-consistent advantage over rate-matched and prefill controls (one-sided sign test $p = 0.016$). (C4) A cross-plant calibration recipe, write amplitude inversely proportional to lens

 doi:10.20944/preprints202607.0895.v1

transport strength, that transfers the method to a second model family with monotone dose response at four seeds, together with a two-plant amplitude law spanning roughly 1.5 orders of magnitude. (C5) Registered negative results that bound the doctrine: gated steering is a legibility mechanism rather than a throughput maximiser, and naive contrastive refusal steering does not reduce attack success on an aligned model. (C6) A head-to-head benchmark against the Jacobian-lens final-target baseline establishing that band-targeted reading is about twice as sensitive at subtle doses and 2.3 times more write-efficient than continuous flooding (Section 6.13). (C7) A recognition-gated jailbreak defence that turns the same activation signal into a deployable moat, matching brute-force attack-success reduction at zero benign over-refusal on models with cleanly separated projections, with an offline predictor that is correct on all four model families tested (Section 6.15). (C8) A reproducible research-loop architecture, expected-free-energy selection over registered menus with dual-closure gates and adversarial review, released in full alongside the code, lens checkpoints, interactive replay applications, and a Model Context Protocol server (Sections 4 and 9). Section 7 draws these together into a statement of the core innovations and their defensibility.

This work extends two prior threads by the author. The companion Symmetry paper demonstrated expected-free-energy selection over interventions for circuit discovery [14]; the present work promotes the same selection principle from choosing interventions inside one experiment to choosing experiments inside a research programme. The refusal-symmetry preprint analysed output-policy capture across jailbreak-style attacks [15]; the negative result reported here on contrastive refusal steering (Section 6.11) is consistent with its conclusion that refusal behaviour on aligned models is not governed by a single easily-steered direction, in line with independent findings [16].

2. Background and Related Work

2.1. Global Workspace Theory and the J-Space Result

Global Workspace Theory holds that cognition is organised around a limited-capacity workspace whose contents are broadcast to, and readable by, many specialist processes [17–20]. Assessments of the theory's applicability to artificial systems have generally proceeded at the level of architectural checklists [21]. The verbalizable-workspace result gave the theory an operational form for transformers: representations that a model can report in words after intervention occupy a small, late-layer subspace (the J-space), identified by the average input-output Jacobian, and this subspace behaves as a functional workspace in the broadcast sense [8]. The companion Jacobian lens transports a residual vector h_ℓ at layer ℓ into final-layer coordinates through $J_\ell = \mathbb{E}[\partial h_{\text{final}} / \partial h_\ell]$ and decodes it with the model's own unembedding [9]. The present work takes both the result and the instrument as given, vendors the reference lens implementation unmodified, and asks the successor question: given a readable and writable workspace, what constitutes disciplined use?

2.2. Activation Steering

Steering by residual-stream addition has an established literature. Activation addition constructs steering vectors from prompt pairs [3]; contrastive activation addition extracts directions from labelled behavioural pairs and applies them at a chosen layer [4]; representation engineering generalises the approach with regression and principal-component readouts [5]; function vectors show that in-context tasks are carried by transportable directions [22]. All of these methods share mechanics with the present work: a direction, a layer, a coefficient. They differ in what this paper argues are the doctrine-level questions. Prior work applies steering either at every position or over a fixed prompt span, verifies success behaviourally, and treats the layer as a hyperparameter. The doctrine tested here adds timing (write only at high-entropy moments), coordinates (write through the band transport $J^\top u$ rather than a raw direction), verification (read the band back through a band-targeted lens rather than inspecting outputs alone), and an explicit intervention budget in nats. The negative result on contrastive refusal steering (Section 6.11) does not contradict the contrastive literature, which reports effects on models and margins where refusal is more weakly consolidated [4,16]; it bounds what a single extracted

 doi:10.20944/preprints202607.0895.v1

direction achieves against an already-aligned 4B instruct model under this paper's registered criteria. The recognition-gated defence of Section 6.15 repurposes the same contrastive signal, extracted by the companion *prayoga* attack library and injected through the doctrine's entropy-gated writer, as an input-conditioned gate rather than a blanket addition, which is what makes the difference between the negative and the deployable defence.

2.3. Lenses and Readback

The logit lens decodes intermediate residual states with the final unembedding [6]; the tuned lens fits per-layer affine probes to correct its distortions [7]. Sparse autoencoders and their successors decompose activations into dictionaries of interpretable features [1,2,23]. The readback protocol used here differs in purpose: the lens is not an offline analysis tool but an online verification instrument inside a control loop, and the sensitivity advantage of band-targeted over final-target reading is quantified head-to-head against the Jacobian-lens baseline (Section 6.13). The band-articulation finding, that a lens targeted at the band recovers content a final-target lens misses at subtle doses, is one of the paper's replicated results (Sections 6.2 and 6.8); the benchmark also delimits it, since the two instruments converge once the write saturates.

2.4. Active Inference for Experiment Selection

Expected free energy balances epistemic value (expected information gain) against pragmatic value (expected preference satisfaction) and is the canonical action-selection quantity of active inference [11–13,24]. The companion circuit-discovery work cast intervention selection inside a single interpretability experiment as a POMDP solved by expected-free-energy minimisation [14]. The present programme lifts the same machinery one level: candidate experiments, not candidate interventions, form the action set; gate verdicts, not attribution observations, form the observation stream; and beliefs replay deterministically from the programme's own committed ledger. Section 4 describes the resulting research-loop architecture and Appendix C gives its parameters.

2.5. Pratyabhij ñ ā

The recognition school (Pratyabhijñā) of Utpaladeva and Abhinavagupta holds that awareness is intrinsically self-aware and that this reflexivity is linguistic: *vimarśa*, the mind's grasp of its own state, is identified with *parā vāk*, the supreme Word (*Īśvarapratyabhijñākārikā* 1.5.13) [10]. The school's analysis distinguishes levels of speech from unarticulated to fully uttered, describes the flash of manifestation before a cognition stabilises (*sphuraṭṭā*) [25], treats received doctrine (*āgama*) as valid only when recognised by the receiving awareness, and catalogues the impairments (*malas*) that restrict a subject's intrinsic freedom (*svātantrya*). Section 3 maps each element to a testable control-engineering clause. The mapping is instrumental: the philosophy contributes hypotheses precise enough to fail, and two of its suggested mechanisms did fail in their initial registered form (Sections 6.4 and 6.11).

3. The Doctrine as Engineering Specification

This section states the five clauses precisely. Throughout, the plant is a frozen decoder transformer with residual stream $h_\ell \in \mathbb{R}^d$ at layer ℓ, unembedding W_U, and next-token distribution p_t at decode step t. The Jacobian lens at layer ℓ is the linear transport $J_\ell = \mathbb{E}[\partial h_{\text{final}} / \partial h_\ell]$, estimated over a generic text corpus, composed with the unembedding to give the read-out $\text{lens}_\ell(h) = W_U J_\ell h$ [9]. Figure 1 summarises the clauses together with the gates that tested them.

 doi:10.20944/preprints202607.0895.v1

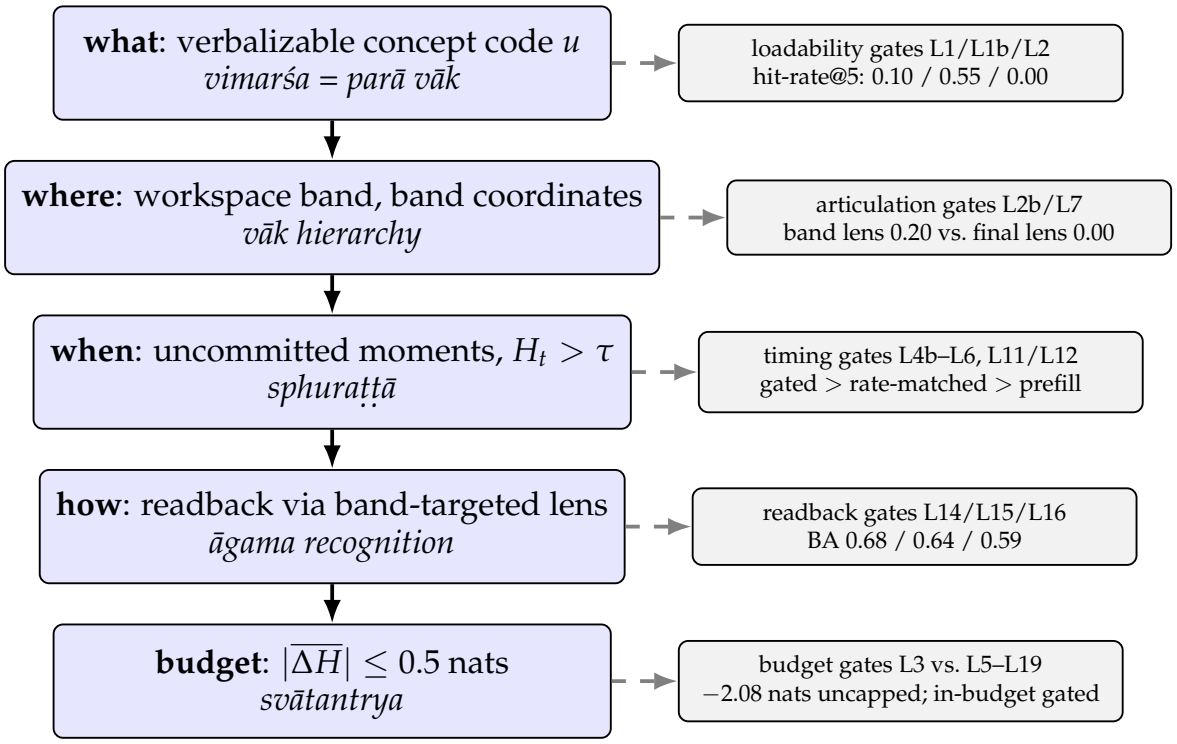

Figure 1. The five doctrine clauses as a pipeline. Each clause (left, with its source concept) is paired with the verification signal and the gate records that tested it (right). Gate identifiers refer to the public ledger (Appendix B).

3.1. What: Verbalizable Concept Codes

A concept code for a nameable concept c (the experiments use fire, metal, river, storm, memory, dream and related stubs) is a unit direction $u_c \in \mathbb{R}^d$ such that adding αu_c to the band residual raises the probability of tokens expressing c. Codes are fitted by linear regression on a small in-distribution corpus and used without tuning on the test corpus. The clause asserts that only verbalizable content loads into the workspace; the loadability experiments (Section 6.1) test whether instructed codes surface at all, per model.

3.2. Where: The Band, in Band Coordinates

The workspace band comprises the late layers (24 to 31 on the 32-layer plants studied) where lens read-outs correlate strongly with model output. Writes target the band and are applied through the transport transpose,

$$h_\ell \leftarrow h_\ell + \alpha\, J_\ell^\top u_c, \tag{1}$$

so that the injected direction is expressed in the coordinates the band itself articulates, rather than in raw residual coordinates. The corresponding failure mode is registered: writing the same code at the wrong depth, or reading the band with a final-target lens, should produce null results. Both predictions are confirmed (Sections 6.2 and 6.7).

3.3. When: Uncommitted Moments

The timing clause holds that writes succeed when the plant has not yet committed to a continuation. Commitment is measured online by next-token entropy $H_t = -\sum_v p_t(v) \log p_t(v)$; a write is permitted at step t only if $H_t > \tau$, with τ self-calibrated per model as an entropy percentile on unsteered traffic. The registered contrasts are a rate-matched control (writes on a fixed schedule at the same average rate) and a prefill control (a single write before decoding). An initial variant that gated on downward entropy spikes, the most literal reading of the flash concept, failed its gate and was revised to the uncommitted-moment form; both the failure and the revision are in the ledger (Section 6.4).

 doi:10.20944/preprints202607.0895.v1

3.4. How to Verify: Band-Targeted Readback

After a write, the verifier reads the band through a lens targeted at the band's own exit and checks the attempted concept against the read-out ranking. The clause encodes the recognition principle that offered content counts as received only when the receiving system re-articulates it. Operationally it makes a sharp, testable prediction: a final-target lens is the wrong verification instrument, and analyses that use it will report false negatives on real band content. Readback quality is quantified as balanced accuracy of hit prediction (Section 6.8).

3.5. Budget: Freedom and Its Cost

Every accepted write perturbs the plant's continuation distribution. The programme prices this in nats: the at-position cost of a write is large and decoding-independent (approximately -2 nats of entropy at the write position), while the trajectory-average cost depends on decoding regime ($+0.82$ nats under greedy decoding, -0.13 under sampling, gates L4 and L4b). The registered budget currency is trajectory-average entropy change under sampling, $\overline{\Delta H}$, capped at ± 0.5 nats. A steering result within this work counts only if it lands inside the cap. The uncapped alternative is documented as a failure case: continuous writing at high amplitude lifted concept surfacing to 0.40 but cost -2.08 nats, violating the budget by a factor of four (gate L3), and the associated failure taxonomy (*malas*: writes rejected for budget, uptake, or quality reasons) is recorded in the same gate.

3.6. Scope

The doctrine governs transparency-oriented steering: making chosen verbalizable content surface in a model's behaviour while preserving its freedom of continuation, with verified uptake. Section 6.11 reports registered experiments on what the doctrine does not provide, namely raw concept-surface throughput and downstream alignment improvements from single contrastive directions.

4. System and Research-Loop Architecture

The programme's second contribution is architectural: the research loop that produced the results is itself a released, replayable system. Figure 2 shows the three tiers.

 doi:10.20944/preprints202607.0895.v1

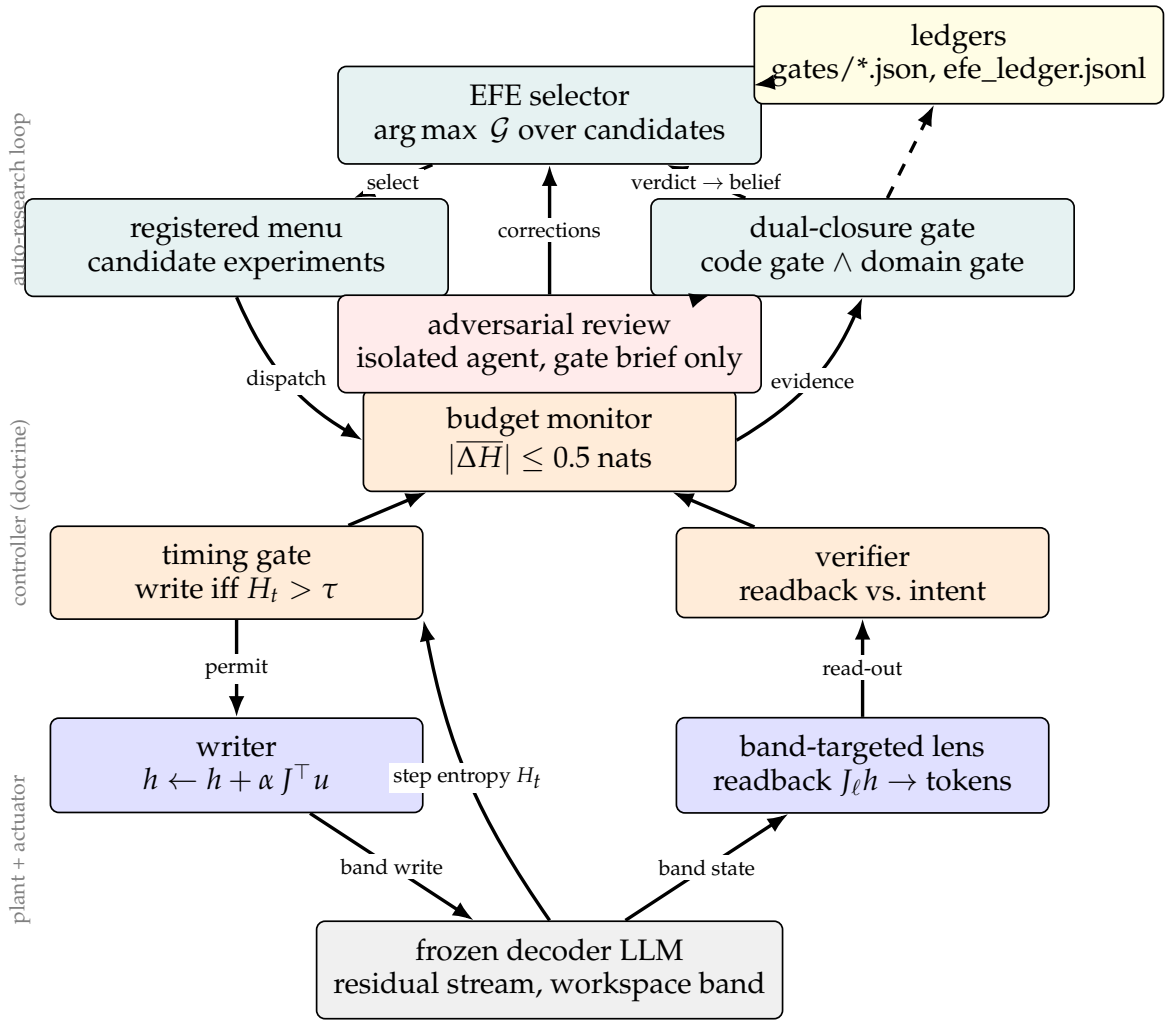

Figure 2. System architecture. Lower tier: the frozen plant with its actuator pair, transported writes (Equation 1) and band-targeted lens readback. Middle tier: the doctrine as controller, an entropy timing gate, a readback verifier, and a budget monitor. Upper tier: the auto-research loop, in which an expected-free-energy selector chooses the next experiment from a registered menu, results close through dual gates, and isolated adversarial reviews feed corrections back. All state flows through committed ledgers.

4.1. Control Loop

At each decode step the controller receives the step entropy H_t, permits a write when $H_t > \tau$ and the budget monitor projects the trajectory-average entropy change to remain inside ± 0.5 nats, and applies Equation 1 at the band. The verifier periodically reads the band through the band-targeted lens and scores the read-out against the intended concept. Timing, amplitude, band membership, and budget are configuration, not code: every tunable lives in a versioned YAML file validated by typed schemas, so a gate record plus a configuration file reproduces a run exactly.

4.2. Dual-Closure Gates

A research loop closes only when two independent gates pass. The code gate requires the test suite and static analysis to be green. The domain gate evaluates the pre-registered scientific criterion for the loop, stated before dispatch with seeds, commands, and thresholds, and issues pass, fail, or fail-on-margin with the raw evidence embedded in the record. The two gates are deliberately non-fungible: several loops in the ledger have passing code gates and failing domain gates (L1, L2, L3, L4, L17, L21), and these closed as registered negative results rather than being re-run to a pass. Each gate record also carries an amendment log; retroactive corrections, such as the sampling-stream repair described below, are visible as amendments rather than silent edits.

 doi:10.20944/preprints202607.0895.v1

4.3. Expected-Free-Energy Experiment Selection

Candidate experiments are registered in menus, each candidate carrying an expected information gain, a preference weight, a compute cost, and a tier (smoke, screen, or confirm, with escalating seed and sample requirements). The selector scores candidates by expected free energy, dispatches the maximiser, observes the resulting gate verdict, and updates its beliefs; beliefs replay deterministically from the ledger of past verdicts, so the entire programme's decision sequence can be re-derived from committed artifacts. Over the programme the selector consumed nine menus in twenty-seven cycles, and its ledger records 267 observation events, 41 proposals, and 27 spend entries. Two formally logged divergences, cases where new evidence superseded an earlier point estimate, were resolved by registered supersession rather than deletion (Appendix C). Contradictions between engineering requirements (steering strength against freedom cost, rigour against velocity, GPU productivity against co-resident job safety) were worked through a TRIZ contradiction log before implementation [26]; the log is released with the repository.

4.4. Adversarial Review

Eighteen reviews were performed by isolated agents given only gate records and contracts, without access to the working tree or the researcher's narrative. Reviews returned verdicts of the form merge-with-corrections, and every correction was applied and committed before the affected loop closed. Material catches include: a cost-model miscalibration in the selector; a ledger-replay bug; the discovery that a single per-run random seed correlated all forty generations within an arm, which invalidated the effective sample size of every earlier sampling estimate and forced per-generation seeding derived from a hash of seed, arm, concept, and stub (the stream fix, gate L9); a candidate name that overclaimed a robustness result (L17); the formal supersession of stale dose-response levels after the stream fix (L18, L19); and the withdrawal of an independence claim that a review showed to be a deterministic replay artifact rather than a replication (L19). Appendix D catalogues all eighteen.

4.5. Compute Discipline

All experiments ran on a single shared DGX Spark node (GB10, unified-memory aarch64) co-resident with unrelated training jobs. A guard script gates every dispatch on GPU idleness, per-loop budget, and a kill-switch file; contention events are recorded in gate metadata. The full programme consumed 18.17 GPU-hours across nineteen budgeted loops (Figure 3), a figure reported to make the cost of the evidence auditable.

 doi:10.20944/preprints202607.0895.v1

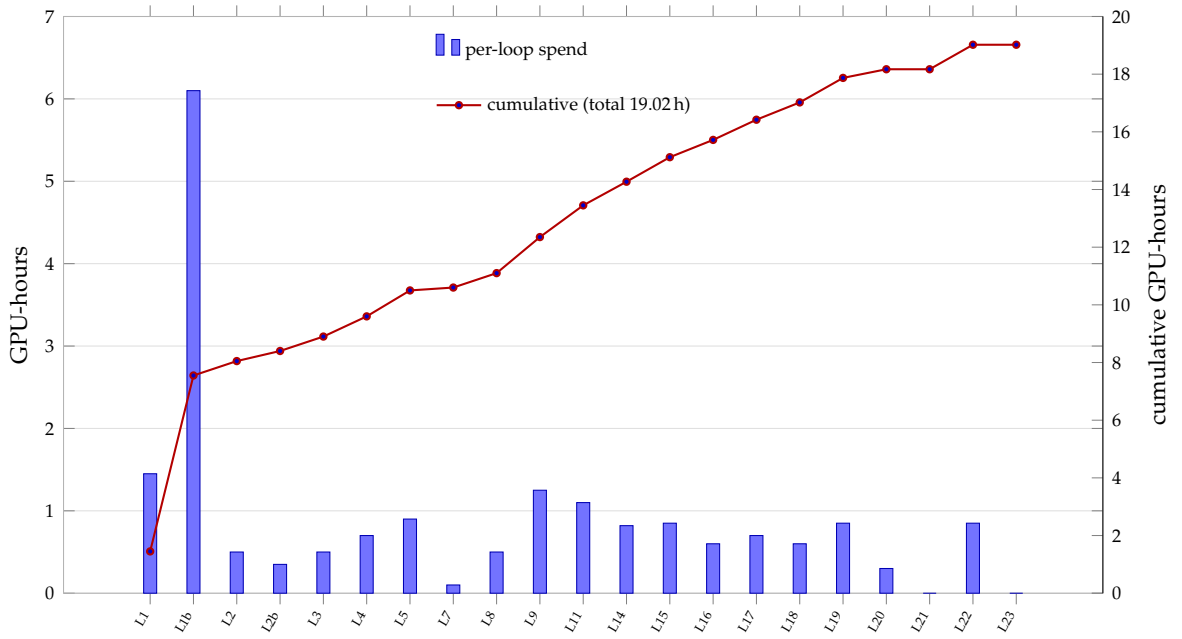

Figure 3. Compute ledger. Bars give per-loop GPU-hours as recorded in the programme state file; the line gives the cumulative total (18.17 GPU-hours). The largest single item is the 27B loadability replication (L1b, 6.1 h).

5. Experimental Setup

5.1. Models

The programme spans five model families and eight distinct checkpoints, all frozen; no weights are updated anywhere. The steering doctrine is developed and confirmed on three plants from two families: Qwen3-4B is the primary plant for cross-plant transfer, the lens benchmark, and comparative evaluation [27]; Qwen3.6-27B, from the same lineage at 6.75 times the parameter count, is used for the loadability scaling comparison; and Nemotron-Mini-4B, a pruned-and-distilled compact model [28] that also serves as the predictive-world-model twin in the author's related stack, is the plant on which the doctrine was first developed. The recognition-gated hardening study (Section 6.15) extends coverage to four further compact instruct models across three additional families: Gemma-2-2B [29], Llama-3.2-1B [30], Qwen2.5-1.5B [31], and SmolLM2-1.7B [32]. Table 1 lists the full set with the loops in which each is used.

Table 1. Model families and checkpoints across the programme. Layer counts and band ranges are the characterised values where measured.

Family	Checkpoint	Size	Role
Qwen3	Qwen3-4B	4B	primary plant, L1–L22
Qwen3	Qwen3.6-27B	27B	loadability scaling, L1b
Nemotron	Nemotron-Mini-4B	4B	doctrine development, twin
Gemma-2	Gemma-2-2B	2B	moat, works (L26)
Llama-3.2	Llama-3.2-1B	1B	moat, works (L26)
Qwen2.5	Qwen2.5-1.5B	1.5B	moat, fails (L26)
SmolLM2	SmolLM2-1.7B	1.7B	moat, backfires (L26)

5.2. Corpora and Concepts

Concept codes are fitted on small in-distribution corpora and evaluated on held-out prompts in three styles: descriptive scenes ($n = 100$ per concept), narrative past ($n = 100$ per concept), and single-token stubs. The primary concept set is fire, metal, river, and storm, extended with memory and dream for the associative-memory experiments. Corpus robustness is evaluated explicitly in loops L16 to L19 rather than assumed.

 doi:10.20944/preprints202607.0895.v1

5.3. Seeding and Independence

Every experiment pre-registers its seeds. Following the L9 stream repair, each generation receives an independent seed derived from a hash of the registered seed, the arm, the concept, and the stub, so arms are compared across genuinely independent sampling streams. Results predating the repair are marked as superseded where later re-runs exist (L18 re-established the dose grid; L19 confirmed the supersession offset at three seeds and two amplitudes). Confirm-tier claims require at least three seeds; the core claim is stated at six.

5.4. Tiers, Gates, and Statistics

Experiments run at three tiers: smoke (functional checks, minutes), screen (single- or few-seed effect estimates), and confirm (multi-seed pre-registered criteria). Significance uses permutation tests against shuffled-code or shuffled-label nulls with 10^4 resamples where distributional assumptions would be doubtful, sign tests across seeds for direction consistency, and registered margins rather than post-hoc thresholds. One early metric was recalibrated after its null was found to be structurally offset: union-top-K rank correlation carries a null floor near -0.72, and all reported loadability figures use model-top-K support, whose null is approximately zero, with permutation gates. Where a gate reports an exploratory sweep (the readback threshold search of L14), the sweep size and the absence of multiple-comparison correction are recorded in the gate and restated here (Section 6.8).

5.5. Reproducibility

Runs execute in a pinned container (Python 3.10, PyTorch 2.4, aarch64 build with flash attention) with configurations mounted from the versioned `configs/experiments/` tree (22 experiment configurations). Every claim in Section 6 cites a gate record in the public ledger by identifier; Appendix B tabulates the complete ledger, and Appendix A documents the gate schema with a verbatim example. Figures are generated by a script that reads only committed gate records, so each plotted coordinate is traceable to a ledger entry (Appendix G).

6. Results

Results are organised by doctrine clause. Every number cites a gate record by identifier; fail verdicts are reported as registered negatives. Findings F1 to F6 confirm mechanisms; F7 to F9 bound scope.

6.1. F1: Loadability Replicates and Scales; Pruning Collapses It

Instructed loading writes a verbalizable concept code into the band and counts the fraction of prompts on which the concept surfaces in the model's top-5 continuation tokens (hit-rate@5, $n = 40$ prompts, shuffled-code null with 200 resamples). On Qwen3-4B the instructed rate is 0.10 against a null of 0.0104 (permutation $p \approx 10^{-4}$, gate L1). On Qwen3.6-27B the rate rises to 0.55 (22 of 40, null 0.068), clearing the registered 0.5 threshold (gate L1b). On Nemotron-Mini-4B, a pruned-and-distilled model, the instructed rate is 0.00, indistinguishable from its null and from an uninstructed control (gate L2). Figure 4 plots all three against their nulls. Loadability is therefore not a generic property of decoder transformers at a given size: it scales with size within a lineage and can be absent at 4B when the training pipeline includes pruning and distillation [28]. Band structure itself is present in all three models (CKA band contrast 0.306, 0.269, and 0.137 respectively, gates L1, L1b, L2), so what varies is writability, not banding. A registered reportability metric behaved anomalously in the same experiments, concentrating on the final layers rather than the late third on both Qwen sizes (Spearman ρ of 0.180 at 4B and 0.124 at 27B against a 0.4 threshold); this is recorded as a failed sub-hypothesis in both gates rather than adjusted away.

 doi:10.20944/preprints202607.0895.v1

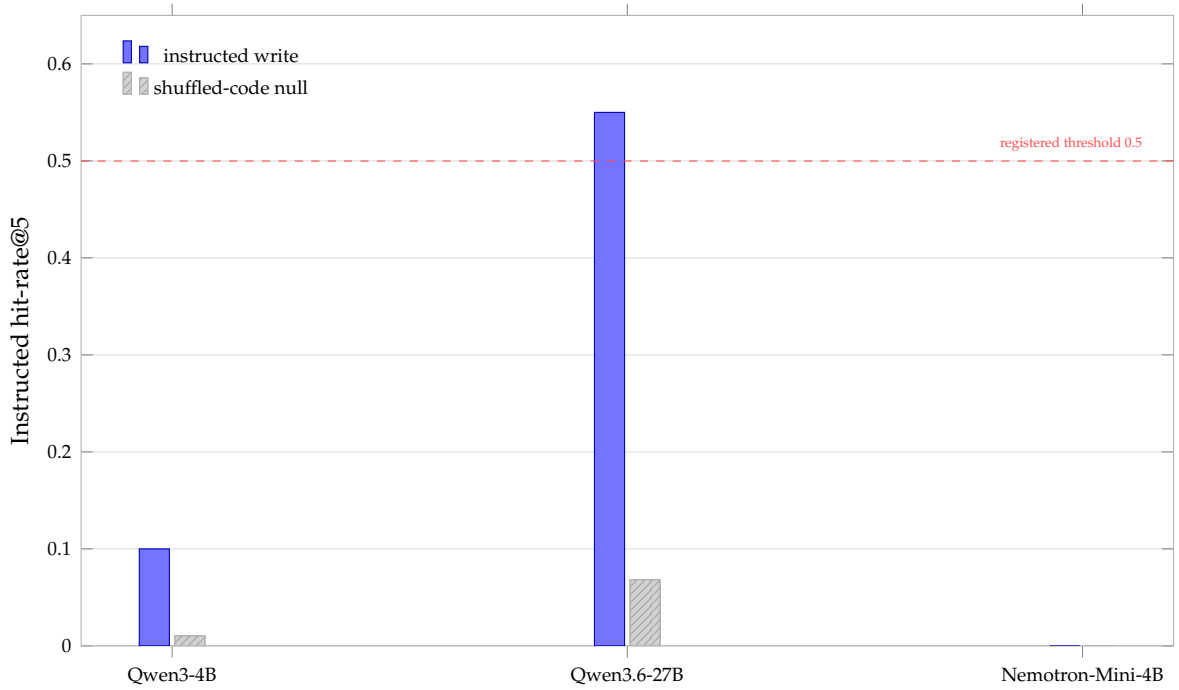

Figure 4. Instructed loadability across plants (gates L1, L1b, L2). Bars give instructed hit-rate@5; hatched bars give the shuffled-code null. The 27B model clears the registered 0.5 threshold; the pruned-and-distilled twin shows no instructed loading at all.

6.2. F2: Band Content Is Legible Only to a Band-Targeted Lens, and the Gradient Is Model-Intrinsic

On the twin whose instructed loading reads as absent, a lens fitted to final output logits reads the band as empty across seven instructed concepts (0.00), while a lens targeted at the band's own exit (layer 26) reads directed loading at 0.20 against a null of 0.023 (gate L2b). The same articulation-depth gradient, read-out negentropy rising through the band, appears on every plant tested. A registered null experiment addresses the concern that the gradient is an artifact of lens construction: the gradient survives under a logit-lens read-out that involves no fitted transport, with Spearman $\rho = 0.607$ against $\rho = 0.639$ for the Jacobian lens, both at permutation $p = 2 \times 10^{-4}$ (gate L7; Figure 5). The gradient is a property of the model's residual stream. The engineering consequence is direct: readback verification must use band-targeted lenses, and final-lens analyses of band interventions produce false negatives.

 doi:10.20944/preprints202607.0895.v1

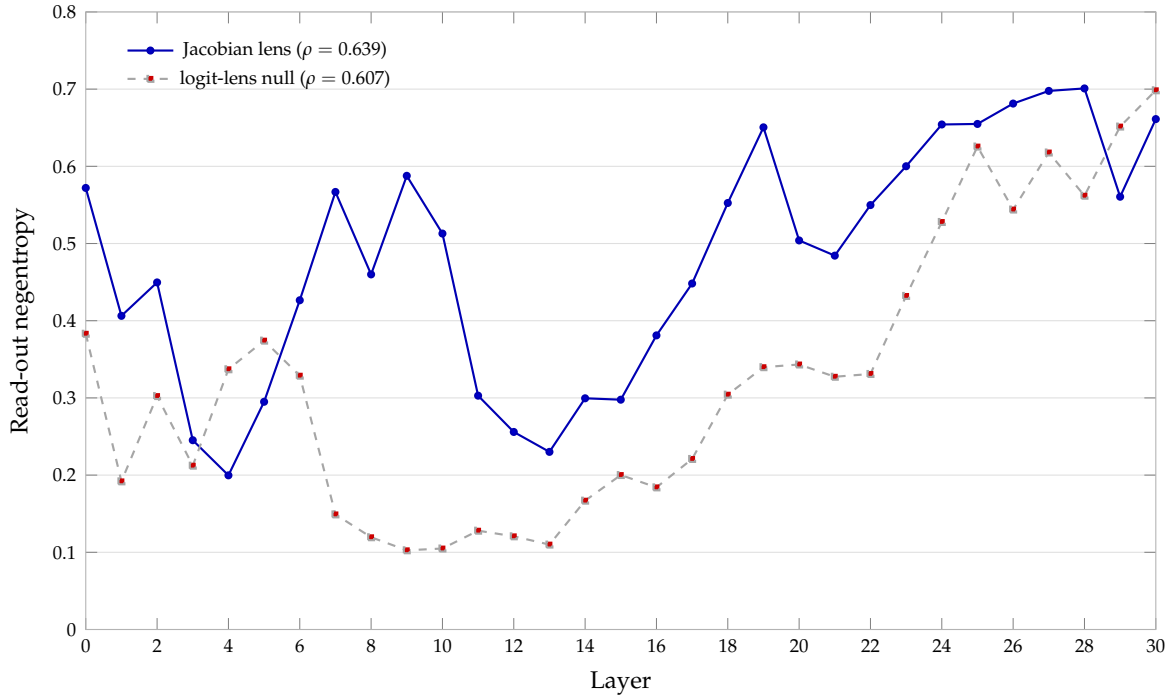

Figure 5. Per-layer read-out negentropy for the fitted Jacobian lens and for a logit-lens null with no fitted transport (gate L7). Both read-outs recover the articulation gradient with permutation $p = 2 \times 10^{-4}$, establishing the gradient as model-intrinsic.

6.3. F3: Uncapped Writing Steers but Destroys Freedom; the Budget Binds

Unrestricted band writing lifts concept surfacing to 0.40 (baseline 0.00) but costs -2.081 nats of trajectory entropy, four times the registered budget (gate L3). The gate's failure taxonomy attributes 20 of the 40 write attempts to freedom-criterion rejection, 14 to uptake failure, and 2 to write-quality failure, identifying freedom cost as the binding constraint. Decomposition across gates L4 and L4b shows the at-position cost of a single write is large and decoding-independent (approximately -2 nats at the write position), while the trajectory-average cost is regime-dependent ($+0.82$ nats greedy, -0.13 sampling); the budget currency was registered accordingly (Section 3.5). These two gates are registered negatives that define the constraint the remaining experiments must satisfy.

6.4. F4: Timing Is the Mechanism; the Flash Variant Failed and Was Revised

The first timing experiment failed its registered criterion: gated lift 0.175 against a 0.2 threshold, with no advantage over schedule or prefill controls (gate L4). Its successor, with corrected gating, passed: entropy-gated writes reach lift 0.40 at trajectory cost -0.127 nats using approximately 9.85 writes per generation, against 0.20 for a single prefill write (gate L4b), the programme's first interventional pass. The threshold τ is not delicate: six of six runs pass across the P40, P60, and P80 entropy percentiles at two seeds, with lift 0.35 to 0.40 and in-budget entropy deltas throughout (gate L5). Timing beats rate: at matched write budgets, entropy-gated placement (lift 0.40 at 7.15 writes per generation) exceeds a rate-matched schedule (0.225 at 5.0) and prefill (0.20), so when to write matters more than how often (gate L6). The literal flash reading of the timing clause, gating on commitment onsets, adds nothing over prefill (0.20 versus 0.30 for uncommitted-moment gating, gate L9); the revised uncommitted-moment form wins at all three seeds tested (advantages of 0.05 to 0.10, gate L12). The doctrine's timing clause survives in revised form, and the revision is documented in the ledger rather than presented as the original hypothesis.

6.5. F5: The Core Claim at Six Independent-Stream Seeds

An adversarial review of the early sampling estimates uncovered that a single per-run seed had correlated all generations within an arm, shrinking the effective sample size (gate L9, the stream

 doi:10.20944/preprints202607.0895.v1

fix; Section 5). All headline claims were therefore restated on independent streams. The core claim, entropy-gated band writes steer within the freedom budget, holds at all six registered seeds, with gated lift 0.30 to 0.35 and trajectory entropy deltas between -0.006 and $+0.246$ nats, all inside the cap (gate L11; Table 2; Figure 6). The advantage over prefill is positive at every seed (mean $+0.100$, range $+0.075$ to $+0.125$), giving a one-sided sign-test $p = 1/2^6 = 0.0156$; its magnitude, however, clears the registered 0.1 margin at only three of six seeds, and an earlier three-seed confirmation had already split the same way (budget claim 3 of 3, margin claim 1 of 3, gate L6 confirm). The claim supported by the evidence is therefore direction, not size: gating always helps, by a margin near 0.1. Steering is also operationally free in throughput terms: steered arms decode at 24.8 to 25.0 tokens per second against a 19.7 baseline on the shared node, so write count imposes no serving cost at this scale (gate L9 write-cost).

Table 2. Core claim at six independent-stream seeds (gate L11). Lift is concept hit-rate@5 over baseline; advantage is gated minus prefill; ΔH is trajectory-average entropy change (budget ± 0.5 nats). Sign consistency 6/6, one-sided $p = 0.0156$.

Seed	Gated lift	Advantage	ΔH (nats)
42	0.30	+0.075	+0.112
123	0.35	+0.125	+0.246
777	0.35	+0.075	+0.109
2024	0.35	+0.125	−0.006
31415	0.35	+0.100	+0.094
999	0.35	+0.100	+0.149
mean	0.342	+0.100	+0.117

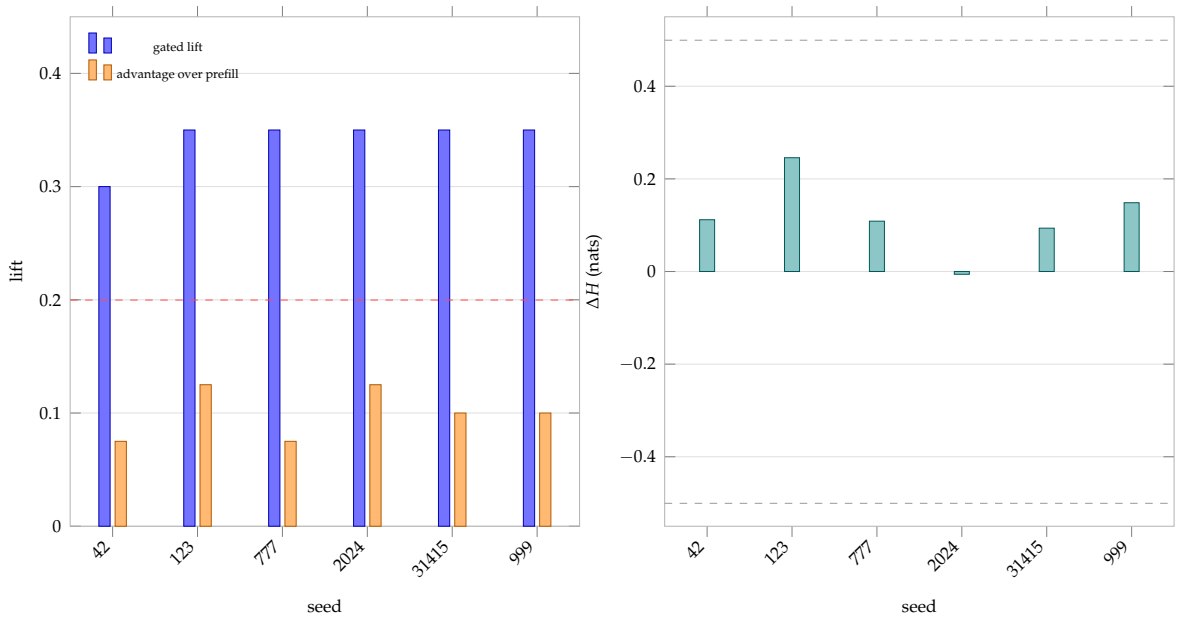

Figure 6. The core claim at six independent-stream seeds (gate L11). Left: gated lift and advantage over prefill per seed; the dashed line marks the registered 0.2 lift threshold. Right: trajectory entropy deltas, all within the ± 0.5-nat budget (dashed lines).

6.6. F6a: Dose Response by Timing Arm, with a Documented Supersession

The canonical dose grid (gate L18, three amplitudes, clean streams) is shown in Figure 7. Continuous writing reaches the highest surface lift (0.45 to 0.50) at 24.7 to 29 writes per generation; entropy-gated writing reaches 0.275 to 0.30 at 8.4 to 9.0 writes; rate-matched and prefill controls sit at 0.15 to 0.225. All gated points are within budget. An earlier version of this grid (gate L8) predated the stream fix and ran approximately 0.1 high at the gated points; the supersession is formal, confirmed at

 doi:10.20944/preprints202607.0895.v1

three seeds and two amplitudes with all six per-seed offsets negative (mean -0.108 at $\alpha = 0.02$ and -0.067 at $\alpha = 0.1$, gate L19), and the superseded gate remains in the ledger with its status marked.

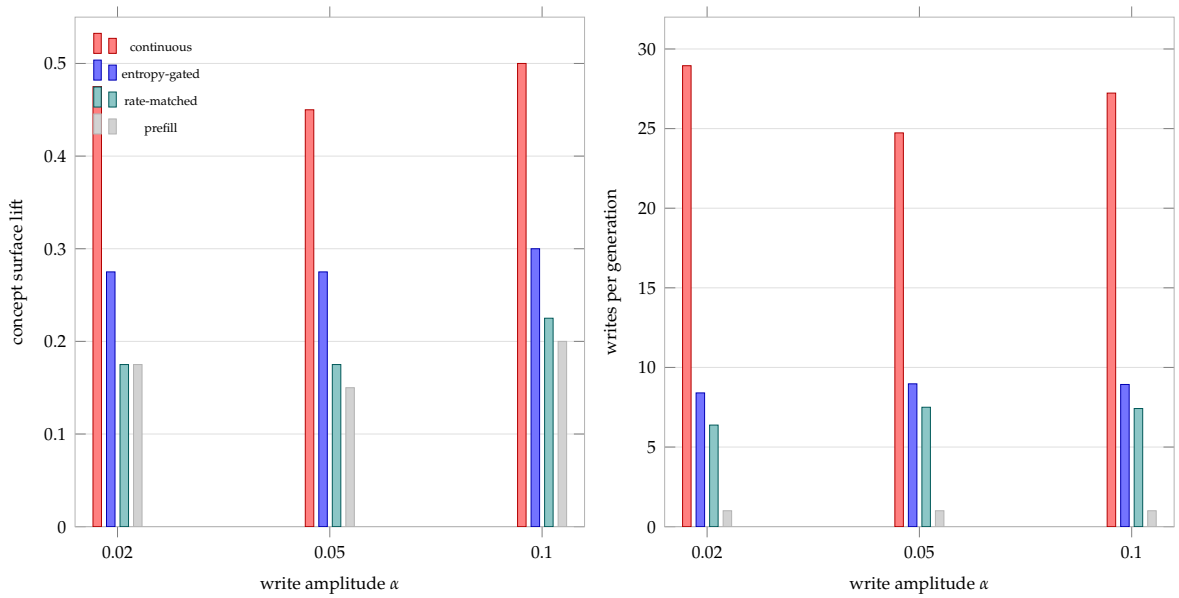

Figure 7. Timing arms across write amplitude on Nemotron-Mini-4B, clean streams (gates L18). Left: concept surface lift; right: writes per generation. Entropy-gated writing obtains roughly 60 % of the continuous arm's lift at a third of its writes, within budget; the continuous arm's raw-lift dominance is quantified further in Section 6.11.

6.7. F6b: The Calibration Recipe Transfers Across Plants; Geometry Does Not

Transplanting the working configuration from Nemotron-Mini-4B to Qwen3-4B unchanged fails: gated lift 0.05, prefill 0.00 (gate L10). The diagnosis is quantitative: the target-layer Jacobian transports on Qwen3-4B are roughly ten times weaker than on the donor, so the same amplitude under-drives the band. Site sweeps eliminate layer choice as the blocker (all probed sites read 0.00 under borrowed amplitude, gate L13 probes). The recipe that transfers sets amplitude inversely proportional to lens transport strength: at three times the donor amplitude, Qwen3-4B reaches gated lift 0.40 against prefill 0.175 (gate L13), the dose curve is strictly monotone over a four-point amplitude grid (0.05, 0.20, 0.40, 0.775 at α from 0.1 to 0.45, gate L14), the recipe holds at four seeds (gated 0.325 to 0.475, prefill 0.075 to 0.175, all in budget, gate L14 multiseed; Figure 8), and per-seed monotonicity replicates at two further seeds (gate L15). On the donor side, a fine amplitude grid finds the same law at one-tenth the scale, rising from 0.025 at $\alpha = 0.002$ to saturation near 0.275 at $\alpha = 0.01$ (gate L16). Together the two plants trace a dose law spanning roughly 1.5 orders of amplitude magnitude (Figure 9). The gate discloses the limits: the two-plant statement is ordering and saturation consistency at screen tier, the within-plant monotone law is the confirmed component, and the donor curve's flat first pair carries no dose signal (gate L15 disclosures).

 doi:10.20944/preprints202607.0895.v1

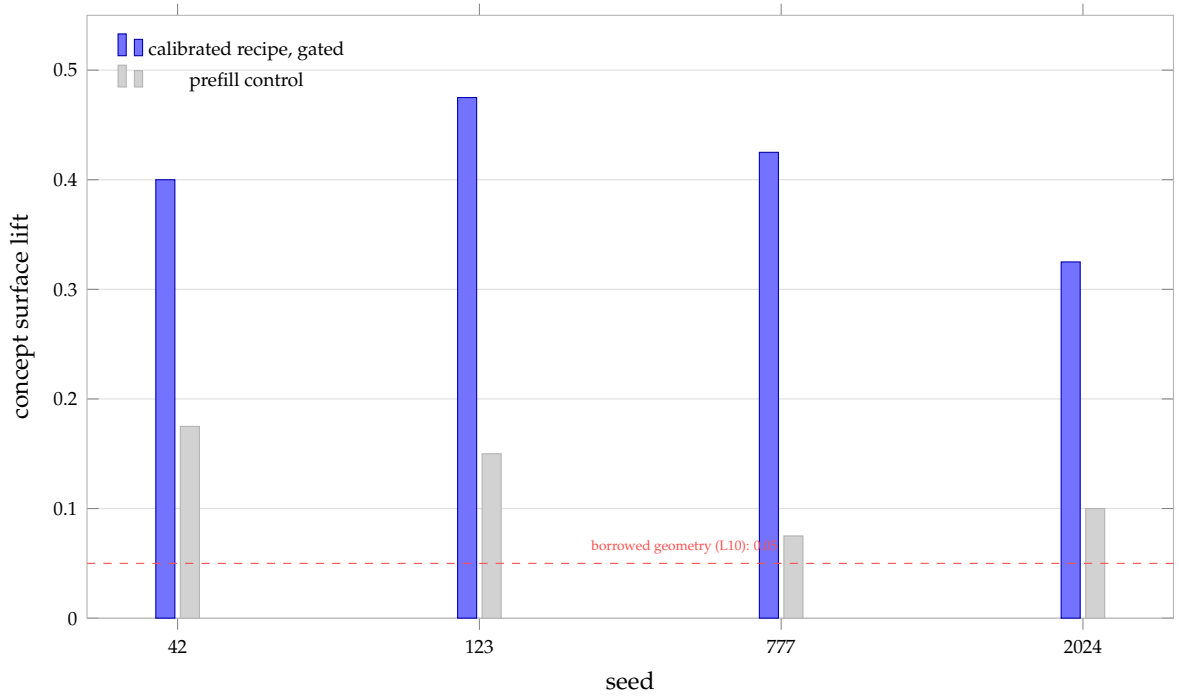

Figure 8. Recipe transfer to Qwen3-4B at four seeds (gate L14 multiseed): calibrated-amplitude gated writes against prefill controls. The dashed line marks the lift obtained by transplanting the donor configuration without calibration (0.05, gate L10).

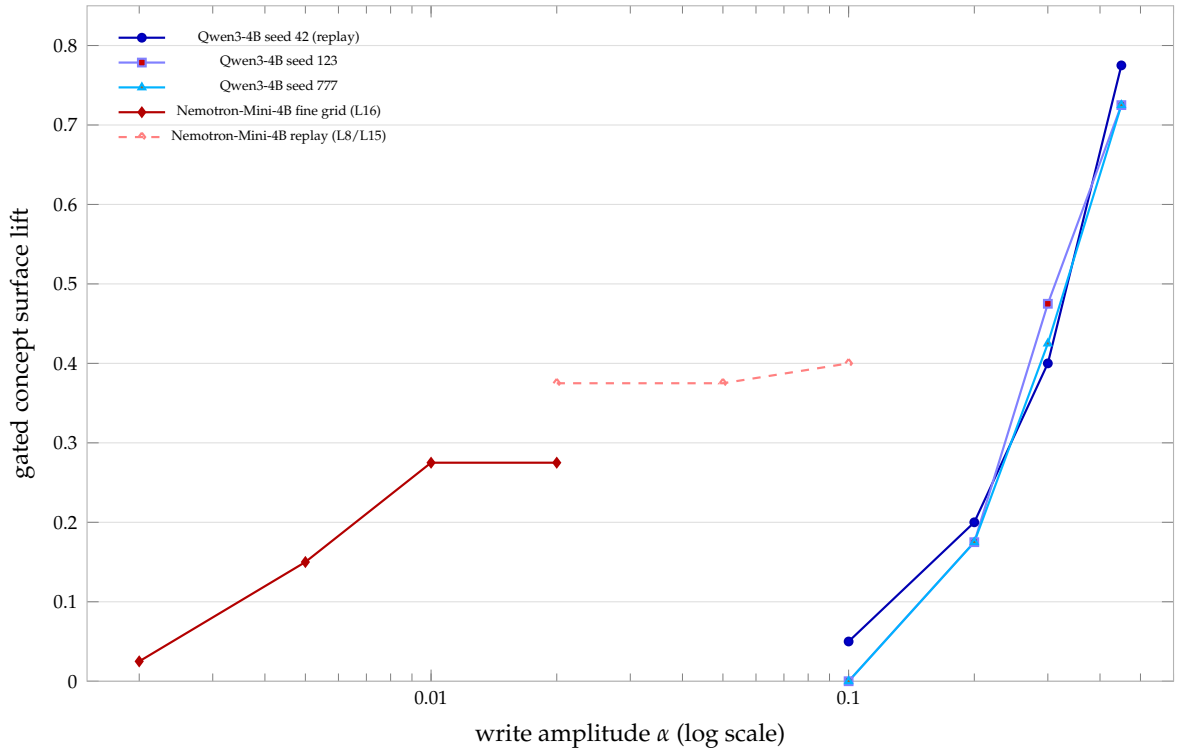

Figure 9. Two-plant amplitude law (gates L14, L15, L16, L8 replay). Gated lift against write amplitude on a log axis: Qwen3-4B at three seeds (blue), Nemotron-Mini-4B fine grid (red, solid) and replayed working points (red, dashed). The working amplitudes differ by an order of magnitude in the direction predicted by inverse lens transport strength.

6.8. F6c: Readback Tracks Behaviour at Screen Tier; Pooled Generalisation Fails

The verification clause requires that band readback predict behavioural uptake. On the development corpus, the best readback setting reaches balanced accuracy 0.684 (true-positive rate 0.833,

 doi:10.20944/preprints202607.0895.v1

true-negative rate 0.536, $n = 40$); the gate disclosal notes that 36 settings were swept, the maximum is reported, and no multiple-comparison correction was applied (gate L14 readback, screen-exploratory). On a held-out corpus with the setting frozen, balanced accuracy is 0.637 against a registered 0.6 threshold, with the margin inside the binomial confidence interval; the same held-out run's steering lift failed its own budget-gate criterion, so lift is corpus-dependent (gate L15 readback, downgraded from confirm to screen by adversarial review). Pooled across three corpus styles ($n = 120$), balanced accuracy falls to 0.590, below threshold, and is recorded as a negative exploratory result (gate L16). The verification clause is therefore supported at screen tier on matched corpora and unproven in pooled generalisation.

6.9. F6d: Corpus Robustness Is an Amplitude Question

Gated steering passes on two of three corpus styles at the working amplitude (narrative-past 0.25, descriptive-scene 0.30, held-out replay 0.15 against a 0.2 bar, gate L16). A seed-variance audit then found the narrative corpus seed-fragile at that amplitude (0.25, 0.075, 0.10 across three seeds) while the descriptive corpus is seed-robust (0.30, 0.20, 0.25), a self-audit fail recorded at 2 of 4 registered cells (gate L17). Doubling the amplitude repairs the fragile corpus at all three seeds (0.425, 0.275, 0.225, one margin thin at 0.225 against 0.2, gate L18), and both corpora approximately double their lift under the amplitude increase (0.175 to 0.325 and 0.275 to 0.55, with one registered per-seed margin failing at 0.025 against 0.05, gate L19). The supported statement is that corpus difficulty loads on the amplitude axis, held at screen tier with one disclosed margin failure; a stronger corpus-amplitude law is registered as open.

6.10. F7: An Untrained Associative-Memory Write Path Matches the Analytic Path

The doctrine's write direction $J^\top u$ is analytic. A biologically-motivated alternative routes the write through recall from a modern-Hopfield associative store [33], the CittaStore of the author's predictive-world-model stack. Integrated end-to-end and run cold, with no patterns stored and no training, the store's recall direction steers within budget on all three seeds (lifts 0.4445, 0.5556, 0.4445, gate L20), and matches the analytic path exactly on two seeds (gaps 0.0000) while differing by a single concept-hit on the third (0.4445 against 0.3334, a gap of 0.1111 against a 0.05 equivalence criterion, in the cold store's favour, at the metric's 1/9 quantisation). The registered equivalence verdict is fail-on-margin at 2 of 3. The store is untrained, so the result is an integration proof that the associative write pathway is live and dimensionally compatible, not evidence that a trained store beats the analytic baseline. Whether training the store improves on cold initialisation is registered as open. Figure 10 plots both arms.

 doi:10.20944/preprints202607.0895.v1

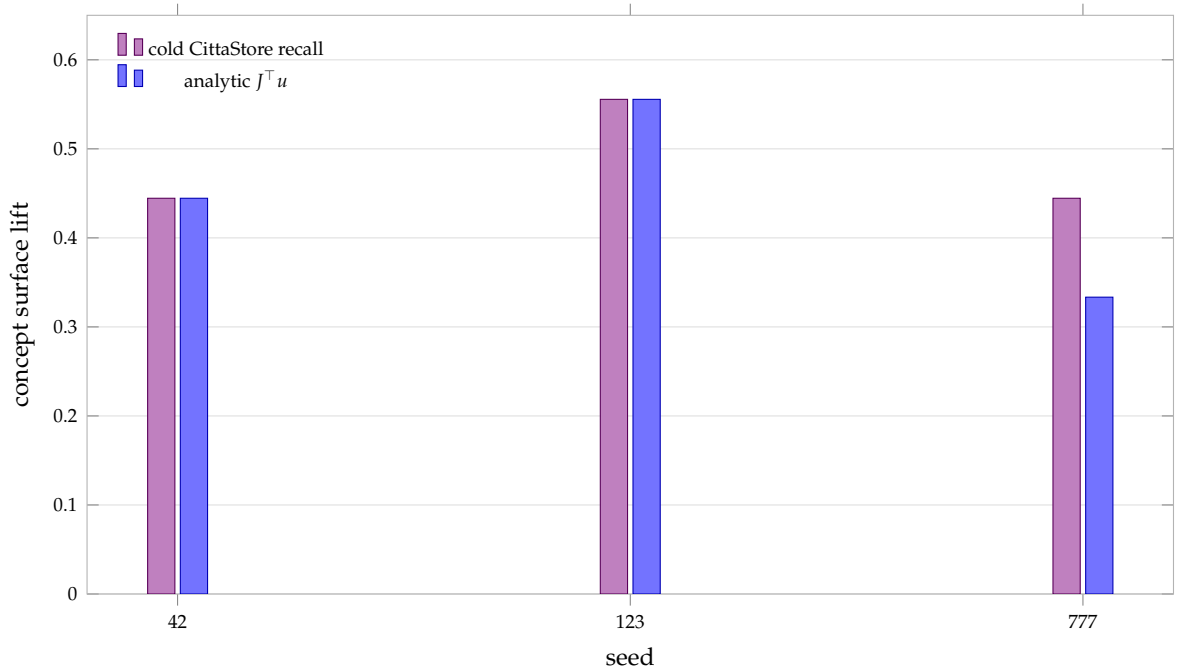

Figure 10. Cold associative-store recall against the analytic write direction at three seeds (gate L20). Two seeds match exactly; the third differs by one concept-hit (the metric's quantisation unit) in the cold store's favour.

6.11. F8: Gated Steering Is a Legibility Mechanism, Not a Throughput Maximiser

A five-arm comparative evaluation on Qwen3-4B ($n = 25$ per arm per seed, three seeds) measures concept surface rate (CSR), a heuristic attack-success proxy (ASR), refusal rate, and entropy cost (gates L21 baselines; Table 3; Figure 11). Continuous flooding surfaces concepts at 0.96; entropy-gated and prefill writes surface at 0.16; a logit-bias arm with no layer or timing structure reaches 0.40. Entropy costs are small in all arms (-0.043 to $+0.073$ nats), so at matched, negligible entropy cost the continuous arm outperforms the gated arm on raw throughput by a factor of six. The gated arm's registered CSR criterion (at least 0.2) fails at all three seeds, and the verdicts are recorded as fails. The reading consistent with the full ledger is that gating purchases verified, budgeted, moment-targeted content placement rather than maximal surfacing; where raw surfacing is the goal, flooding a 4B model is easy and needs no doctrine. This reframes the doctrine's value as internal-model transparency under constraint, and Section 8 develops the point.

Table 3. Comparative evaluation on Qwen3-4B (gates L21 baselines, mean of seeds 42, 123, 777; $n = 25$ per arm per seed). CSR is concept surface rate; ASR is a heuristic attack-success proxy on the same prompt set; ΔH is mean entropy cost in nats. Per-seed values are identical to three decimals across seeds.

Arm	CSR	ASR	Refusal	ΔH
baseline	0.000	0.800	0.200	$+0.000$
prefill	0.160	1.000	0.000	-0.020
entropy-gated	0.160	1.000	0.000	-0.020
logit bias	0.400	0.800	0.200	-0.043
continuous	0.960	1.000	0.000	$+0.073$

 doi:10.20944/preprints202607.0895.v1

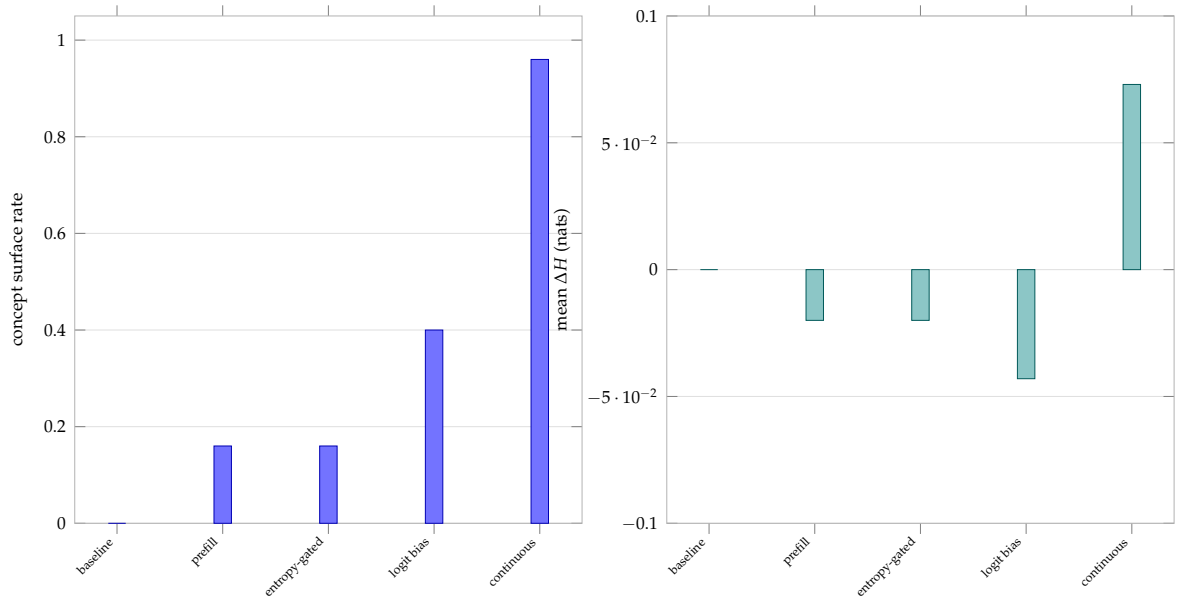

Figure 11. Comparative evaluation arms (gates L21 baselines, three-seed means). Left: concept surface rate; right: mean entropy cost. Continuous flooding dominates raw surfacing at negligible entropy cost; the gated arm's value is verified placement, not throughput.

6.12. F9: Naive Contrastive Directions Do Not Improve Alignment on an Aligned Model

Two registered experiments test whether contrastive activation directions [4,5] improve safety-relevant behaviour on the already-aligned Qwen3-4B instruct model. A refusal direction extracted from labelled harmful and benign pairs and applied at the working site leaves attack success on an AdvBench subset [34] unchanged: ASR 0.250 with and without steering, refusal rate 0.750 in both arms ($n = 20$, gate L21 jailbreak, registered fail). A truthfulness direction extracted from TruthfulQA contrasts [35] fails its registered improvement criterion, with the gate recording identical heuristic scores in both arms ($n = 15$, gate L21 truthful, registered fail). The gates disclose their limits: heuristic pattern-based metrics rather than judge models, single-seed direction extraction, and small subsets. Within those limits the result is consistent with the direction-structure analysis of the author's refusal-symmetry work [15] and with evidence that refusal behaviour, while direction-mediated, is not improved by naive single-direction addition on models whose alignment margin is already consolidated [16]. Alignment gains at the activation level are not an automatic dividend of naive workspace steering. Finding F12 shows that the same activation signal, used for input-conditioned recognition rather than blanket addition, does provide a usable defence on some models.

6.13. F10: Band-Targeted Reading Is Twice as Sensitive as the Final-Target Baseline at Subtle Doses

A head-to-head benchmark compares the paper's band-targeted lens against the Jacobian-lens final-target baseline as reading instruments applied to the same deterministic writes, on a fixed grid of ten concepts by eight stubs ($n = 80$ paired observations, exact McNemar test, since a deterministic forward pass cannot be reseeded; gate L22, consolidated from committed sub-gates by a recomputation script). Across a write-dose floor sweep (Figure 12) the two instruments are indistinguishable at doses too small to load the band (both 0.00 at $\alpha = 0.02$, both 0.025 at $\alpha = 0.05$), then separate sharply at $\alpha = 0.1$, where the band-targeted lens detects on 0.475 of pairs against 0.2375 for the final-target lens, a gap of $+0.2375$ with 20 band-only detections against one final-only ($p = 2.1 \times 10^{-5}$). At the saturating dose $\alpha = 0.3$ both instruments approach ceiling (band 1.00, final 0.95, gap $+0.05$, $p = 0.125$). The registered gap criterion of at least 0.3 is missed by 0.0625, so the verdict is fail-on-margin, and the strong-form necessity claim (that band content is wholly invisible to a final-target lens) is withdrawn under its own falsifier, since the final-target lens does recover most content once the write saturates. The supported statement is that band-targeted reading is roughly twice as sensitive at subtle, budget-

 doi:10.20944/preprints202607.0895.v1

respecting doses, with the direction decisive at $p = 2.1 \times 10^{-5}$, and that the two instruments converge only when the intervention is driven hard enough to be detectable by any reasonable readout.

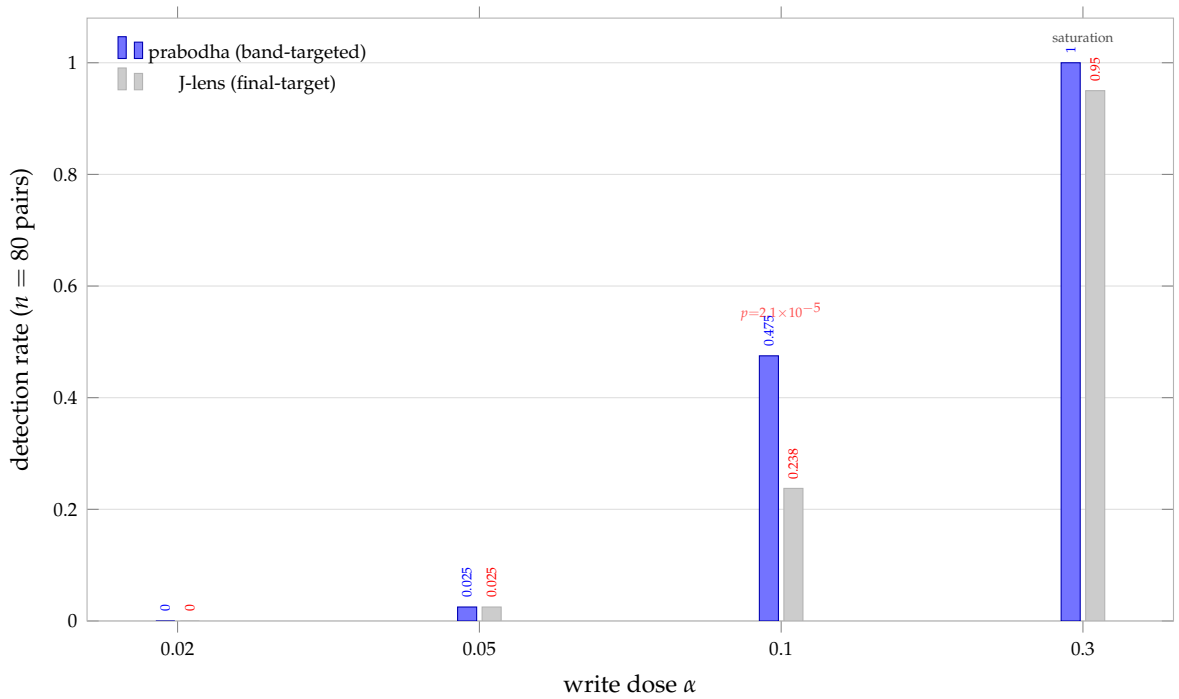

Figure 12. Band-targeted lens (this work) against the Jacobian-lens final-target baseline, detection rate across the write-dose floor sweep (gate L22, $n = 80$ pairs). The instruments diverge at the subtle dose $\alpha = 0.1$ ($p = 2.1 \times 10^{-5}$) and converge at saturation.

6.14. F11: Gated Placement Is 2.3 Times More Write-Efficient than Continuous Flooding

The same benchmark quantifies the efficiency of gated placement against continuous flooding in lift per write, across six clean-stream cells (seeds 42, 123, 777 by amplitudes 0.02 and 0.1, sourced from the superseded-corrected L18 and L19 gates). Gated lift per write exceeds continuous in every cell (sign consistency 6 of 6), with a mean ratio of 2.32 and a range of 1.83 to 3.25 (Figure 13). Expressed in absolute terms, gated placement recovers 67 percent of the continuous arm's lift while spending 29 percent of its writes. Finding F8 established that continuous flooding wins on raw surfacing; Finding F11 establishes that when the cost of a write is counted, gated placement is the more efficient instrument, which is the regime in which the doctrine's timing clause pays for itself.

 doi:10.20944/preprints202607.0895.v1

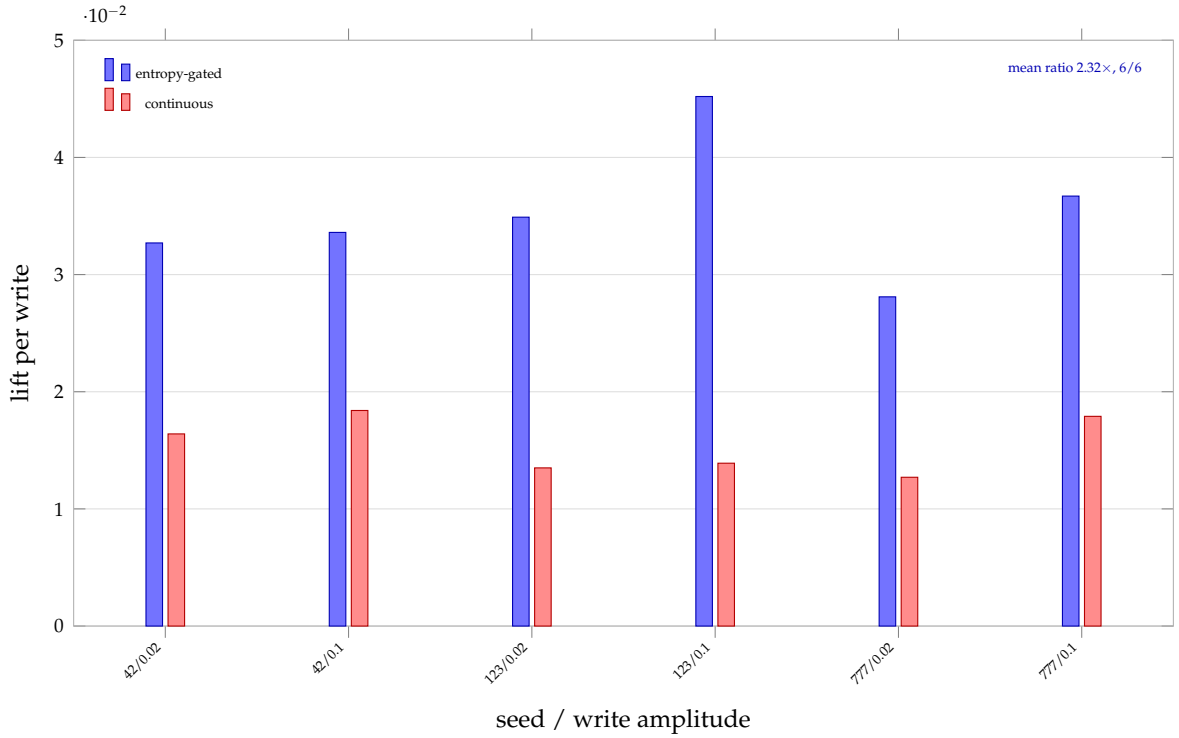

Figure 13. Lift per write, entropy-gated against continuous, across six seed-by-amplitude cells (gate L22 efficiency block). Gated placement wins every cell; mean ratio 2.32, recovering roughly two-thirds of the lift at under one-third of the writes.

6.15. F12: Recognition-Gated Hardening Is a Deployable Moat on Some Models, Predicted Before Deployment

The activation signal that fails as a blanket steering direction (F9) succeeds when it is used the way the doctrine's verification clause suggests, as a recognition test on the input rather than an unconditional write. A defence built on the fused attacker-defender loop (external refusal-direction extraction supplied by the companion *prayoga* attack library, injected through prabodha's entropy-gated writer and, in the final form, conditioned on an input-projection recognition gate) hardens a model against jailbreaks only when the input's projection onto the refusal direction crosses a threshold placed in a clean gap between benign and attack projections. On Gemma-2-2B the benign and attack projection ranges are cleanly separated at read layer 13 (benign $[-53, -18]$, attack $[+4, +73]$); placing the threshold in that gap cuts real jailbreak attack success from 0.50 to 0.25, matching brute-force unconditional hardening, but at zero benign over-refusal, whereas the unconditional defence reaches the same 0.25 only by refusing every benign prompt (over-refusal 1.00), which is unusable (gate L26 proof, exploratory, $n = 12$ attacks and 10 benign, single seed).

Generality is bounded and, importantly, predictable. Across four model families the moat works on exactly two, and the predictor is the clean-gap test computed before any deployment (Figure 14, Figure 15; Table 4). Gemma-2-2B (0.50 to 0.25) and Llama-3.2-1B (0.25 to 0.083) have cleanly separated projection ranges and harden at zero benign over-refusal. Qwen2.5-1.5B and SmolLM2-1.7B have overlapping ranges: on the former the gate cannot discriminate and attack success is unchanged while over-refusal rises, and on the latter the underlying difference-in-means direction is not a clean refusal direction, so unconditional hardening raises attack success from 0.583 to 0.917. The operational consequence is that the same read-layer projection statistic that decides whether to fire the gate at inference time also predicts, offline and cheaply, whether the moat will hold for a given model, so a deployer can test before trusting. These results are exploratory (small samples, single seed, refusal-phrase attack metric), and they are reported as an existence-and-predictability result rather than a general guarantee.

 doi:10.20944/preprints202607.0895.v1

Table 4. Recognition-gated hardening across four model families (moat consolidation, from the released `moat_models.json`). ASR is jailbreak attack success rate; OR is benign over-refusal. The clean-gap column is the offline predictor.

Model	Clean gap	ASR none	Recognition-gated ASR	Recognition-gated OR	Moat
Gemma-2-2B	yes	0.500	0.250	0.00	works
Llama-3.2-1B	yes	0.250	0.083	0.00	works
Qwen2.5-1.5B	no	0.333	0.333	0.20	fails
SmolLM2-1.7B	no	0.583	0.917	0.00	backfires

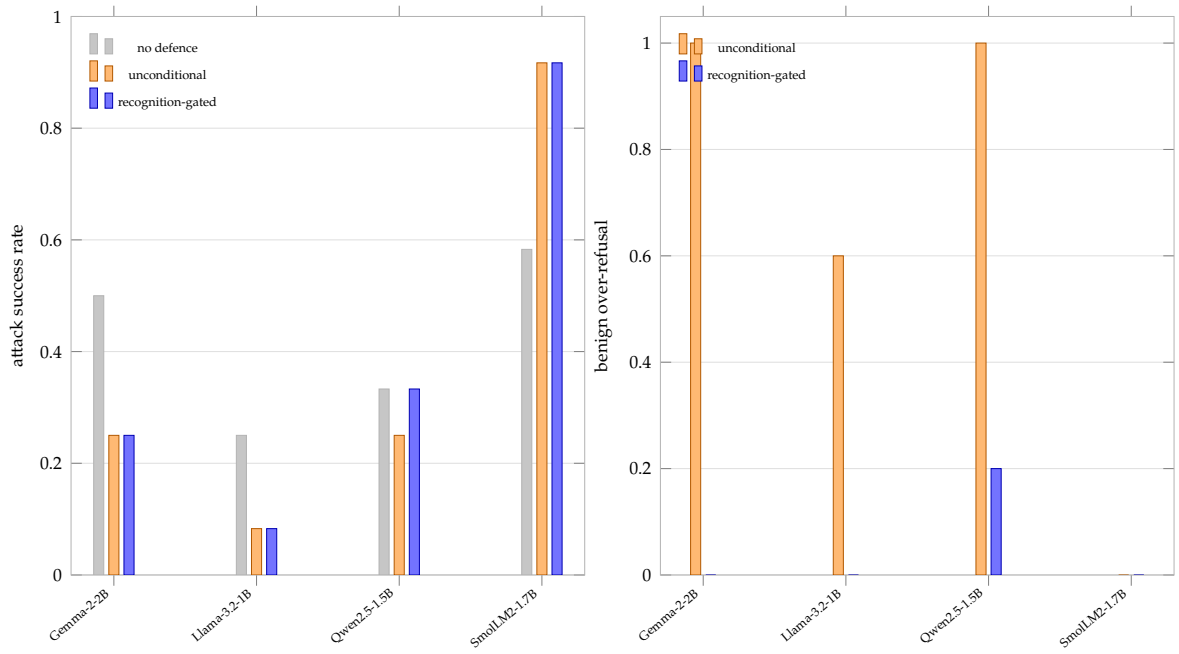

Figure 14. Four-model recognition-gated hardening (moat consolidation). Left: jailbreak attack success under no defence, unconditional hardening, and recognition-gated hardening. Right: benign over-refusal. Recognition gating matches the unconditional attack-success reduction where the moat holds, at a fraction of the over-refusal cost.

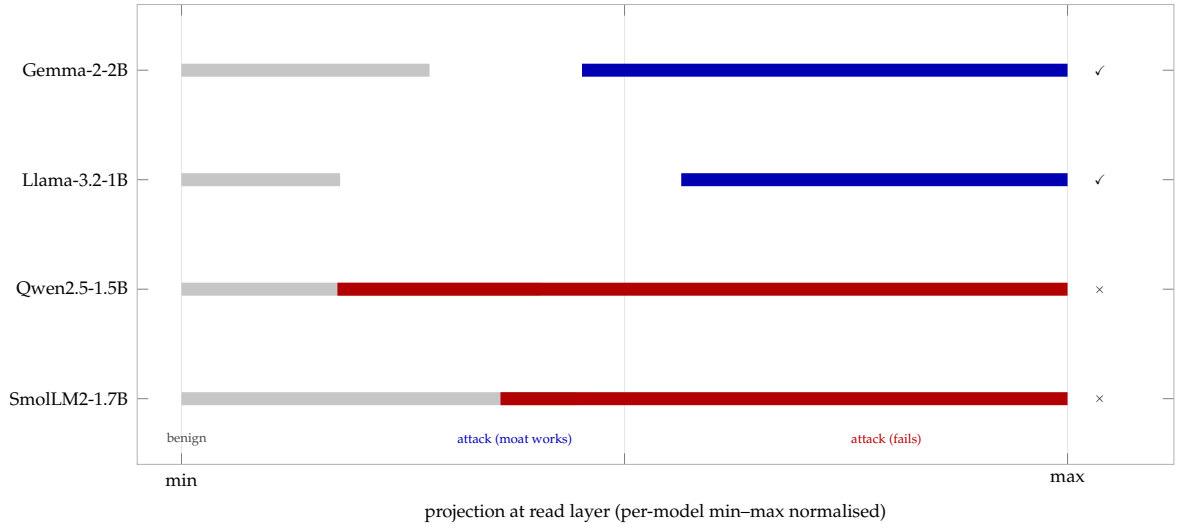

Figure 15. The clean-gap predictor. Benign (grey) and attack (coloured) projection ranges at each model's read layer, normalised per model. Separation predicts the moat: Gemma-2-2B and Llama-3.2-1B separate cleanly and harden; Qwen2.5-1.5B and SmolLM2-1.7B overlap and fail.

 doi:10.20944/preprints202607.0895.v1

The hardening loop was also characterised at the activation level directly. Both Qwen3-4B and Nemotron-Mini-4B are prompt-robust but activation-vulnerable: prompt-only jailbreaks fail entirely (attack success 0.00) while activation-level refusal ablation succeeds (0.90 on Qwen3-4B, 1.00 on Nemotron-Mini-4B), which locates the harmful signature in the residual stream rather than the prompt and motivates an activation-level, rather than prompt-level, defence (gates `char`). An earlier form of the harden loop that added the direction unconditionally under entropy gating preserved freedom and cut writes (11.8 against 50 per generation) but did not resist an activation-level attacker (gate L23, registered fail), which is precisely the negative that drove the design toward input-conditioned recognition (Figure 16).

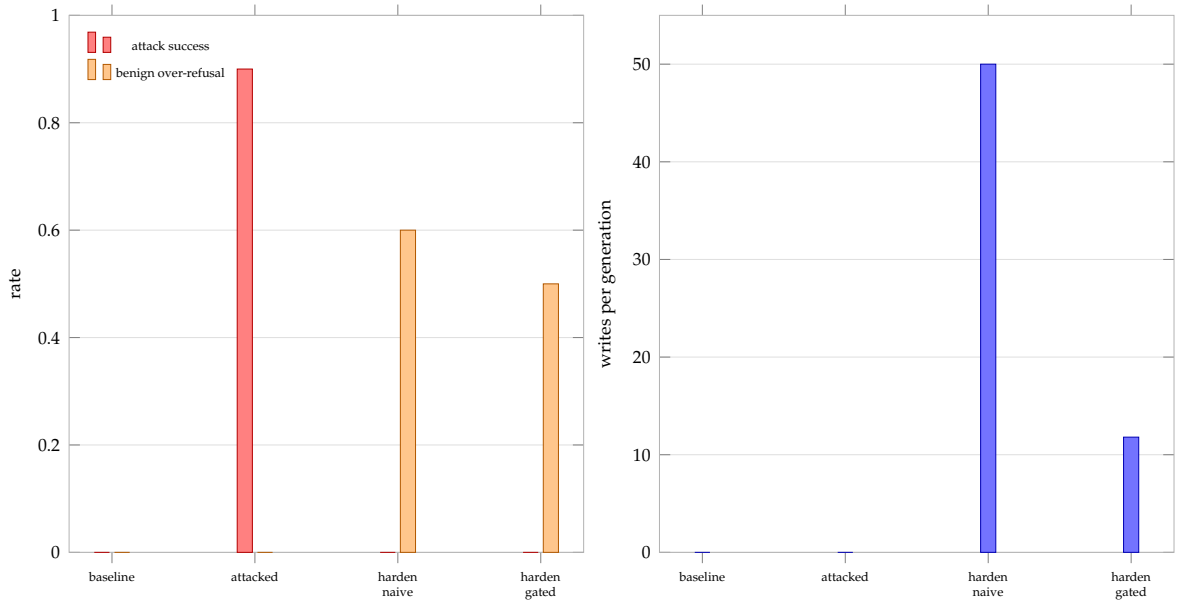

Figure 16. The fused attacker-defender harden loop (gate L23). Left: attack success and benign over-refusal per arm; entropy-gated hardening matches naive hardening on benign generations while using far fewer writes (right), but neither resists an activation-level attacker, motivating the recognition gate of F12.

7. Core Innovations and Defensibility

The programme's contribution is not a single steering trick but a coherent stack in which each layer is independently defensible and, taken together, hard to reconstruct from any one published component. This section states the core innovations explicitly and identifies what distinguishes them from the closest prior art.

7.1. a Doctrine, Not a Vector

The dominant mode in activation steering is to publish a direction and a layer. The innovation here is a doctrine that governs the direction: where to write in the band's own coordinates, when to write by online entropy, how to verify by band-targeted readback, and how much to spend against a registered freedom budget. Each clause is separately falsifiable and two failed in their first registered form (Sections 6.4 and 6.15), which is the evidence that the clauses carry content rather than restating a convention. The measured ablations quantify what each clause buys: removing coordinate discipline collapses transfer lift from 0.40 to 0.05, removing timing costs roughly half the gated advantage, removing the budget quadruples the entropy cost, and using the final-target lens for verification misreads a loaded band as empty at subtle doses.

7.2. Reading in the Band's Own Basis

The second innovation is instrumental. Prior lenses read intermediate states toward the final output basis [6,7]; this work reads and writes in the coordinates the workspace band itself articulates, through the transport transpose $J^{\top}u$. The head-to-head benchmark against the Jacobian-lens final-

 doi:10.20944/preprints202607.0895.v1

target baseline (Section 6.13) shows the practical payoff: at the subtle, budget-respecting doses where an intervention should operate, band-targeted reading is roughly twice as sensitive, with the direction decisive at $p = 2.1 \times 10^{-5}$. The honest bound is equally part of the innovation: at saturating doses the instruments converge, so the claim is sensitivity at low dose, not universal invisibility.

7.3. Recognition as a Deployable Defence

The third and most consequential innovation is the recognition gate. The finding that a contrastive direction added unconditionally does not improve safety (Section 6.12) is, on its own, a negative shared with prior work. The advance is to use the same activation signal as a recognition test on the input rather than as a blanket write, and to show that this converts an unusable brute-force defence (which over-refuses every benign prompt) into one that matches the brute-force attack-success reduction at zero benign collateral on models whose benign and attack projections separate cleanly (Section 6.15). Two properties make this a defensible moat rather than a demonstration. First, it operates at the activation level, so prompt-rewrapping attacks that defeat prompt-level filters do not evade it, consistent with the characterisation that these models are prompt-robust but activation-vulnerable. Second, and unusually for a safety intervention, its success is predictable before deployment from the same clean-gap statistic that drives it at inference time, so a deployer can determine offline and cheaply whether the defence will hold for a given model. The four-model result, works on two and predicted on all four, is presented as an existence-and-predictability claim under exploratory conditions, not a universal guarantee.

7.4. a Reproducible Research Instrument

The fourth innovation is methodological and is released as running software. The expected-free-energy selector over registered menus, closed by dual code-and-domain gates and audited by isolated adversarial reviewers, produced a programme in which every decision replays from committed artifacts and every reported number traces to a gate record. The audit trail did load-bearing work rather than decoration: it caught a sampling-stream correlation that re-based every early estimate, forced a formal supersession of stale dose levels, and withdrew a claimed replication that proved to be a determinism artifact. This instrument, together with the companion circuit-discovery agent [14], constitutes a template for small-compute interpretability research whose claims survive adversarial scrutiny.

7.5. an Integrated, Released Product

The innovations are not left as a paper. They ship as a Python package, fitted lens checkpoints for the studied models, a Model Context Protocol server and agent plugin that expose the lens-and-steer loop as callable tools, and interactive applications that render internal-state traces and replay the recognition-gated defence. Figure 17 shows the paper-static analogue of the interactive workspace view: a layer-by-token map of lens read-out grounded in the measured articulation gradient, annotated with a gated write and its band-targeted readback, in the spirit of the slice visualisation released with the workspace instrument [8,9]. The defensibility of the whole rests on the combination: the doctrine, the band-basis instrument, the predictable recognition gate, the audited research loop, and the released tooling reinforce one another, and reproducing the stack requires all of them rather than any single vector or lens.

 doi:10.20944/preprints202607.0895.v1

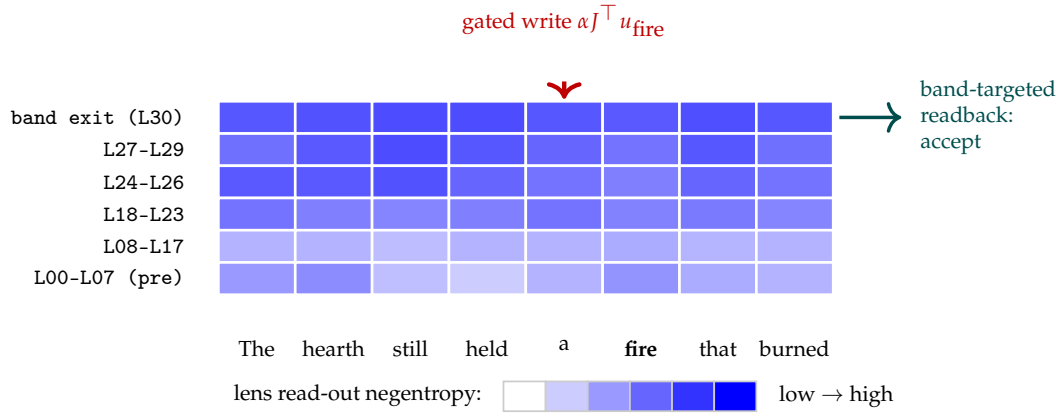

Figure 17. Workspace read-out for a gated fire-case decode, the paper-static analogue of the interactive trace player in the released application. Cell shading is lens read-out negentropy by layer band and decode position, grounded in the measured articulation gradient (gate L7); the red arrow marks the entropy-gated write of the fire concept code at an uncommitted moment, and the readout at the band exit records the band-targeted acceptance verdict. This is the internal-state view that the doctrine reads and writes.

8. Discussion

8.1. What the Doctrine Buys

The confirmed clauses compose into a working control loop: verbalizable codes written through the band transport, at uncommitted moments, verified by band-targeted readback, inside a registered freedom budget. Each clause earns its place by a measured contrast. Removing coordinate discipline (writing borrowed geometry on a new plant) collapses lift from 0.40 to 0.05; removing timing discipline (rate-matched schedules) costs roughly half the gated advantage; removing the budget (continuous high-amplitude writing) quadruples the entropy cost; using the wrong verification instrument (a final-target lens) misreads loaded bands as empty. The doctrine is thus not a bundle of conventions but a set of independently falsifiable constraints, two of which failed in their first registered form and were revised under audit.

8.2. Transparency Rather than Throughput

The comparative evaluation forces a precise statement of what gated steering is for. If the goal is making a 4B model mention fire, continuous flooding achieves 0.96 surface rate at negligible entropy cost and requires no theory. The gated arm's 0.16 at the same cost would be a poor trade were throughput the objective. The objective the doctrine actually serves is different: placing chosen, verbalizable content into the model's workspace at moments when the model is receptive, verifying uptake through the model's own articulation, and bounding the intervention's effect on the model's continuation freedom. These properties, placement, verification, and budget, are what an auditor or a safety pipeline needs from a steering primitive, and none of them is provided by flooding. The finding also cautions against reading raw steering lift as an alignment-relevant capability measure.

8.3. Loadability as a Model Property

The loadability gradient (0.10 at 4B, 0.55 at 27B, 0.00 on the pruned twin) suggests that instructed workspace loading is an emergent capability with a size threshold inside a lineage, and that compression pipelines can remove it entirely while leaving band structure intact. The practical consequence runs in both directions: workspace steering of the kind studied here may become easier on larger models, and distillation may be a deliberate defence against it. Both directions are testable and neither is claimed beyond the three plants measured.

 doi:10.20944/preprints202607.0895.v1

8.4. the Research Loop as a Result

The programme's methodological finding is that an expected-free-energy selector over registered menus, closed by dual gates and audited by isolated reviews, can run a multi-week interpretability programme with every decision replayable from committed artifacts. The audit trail did real work: the stream fix re-based every sampling estimate in the programme, the L18 supersession corrected stale dose levels by a measured -0.10, and a claimed replication was withdrawn when review showed it to be a deterministic replay artifact. These corrections are part of the released record, and the resulting ledger, loops L0 to L21 with twenty-seven selector cycles and eighteen reviews, is offered as a reusable template for small-compute rigorous interpretability work, complementing the intervention-level agent of the companion paper [14].

8.5. Relation to the Source Philosophy

The engineering reading extracted from Pratyabhijñā proved productive in a specific, limited sense: it generated a clause structure whose elements were independently falsifiable, two of which failed as first registered (the commitment-flash timing variant and any expectation of alignment dividends) and were revised or bounded accordingly. Nothing in the results bears on the philosophy's own claims about consciousness, and the paper asserts no such connection. The Sanskrit terminology is retained as precise internal nomenclature, in the same spirit as the companion works' usage [14,15], with a complete engineering glossary in Appendix E.

8.6. Limitations

Model scale and coverage are the primary limits: interventional results are established at 4B on two families, with 27B evidence restricted to loadability, so extrapolation of steering behaviour to larger models is unsupported. Behavioural metrics in the comparative evaluation are heuristic and pattern-based; judge-model scoring would strengthen the alignment-facing conclusions, and this limit is registered in the gates themselves. The contrastive experiments test single directions at a single site and single extraction seed, bounding naive application only; richer contrastive protocols could behave differently. Readback verification holds at screen tier on matched corpora and failed pooled generalisation, so the verification clause is the weakest confirmed link. Concept coverage is narrow (a handful of concrete nouns), corpus styles number three, and the corpus-amplitude interaction is supported at screen tier with one disclosed margin failure. The associative-store result is an integration proof under a cold store, and the trained-store comparison remains open. Finally, the compute envelope (18.17 GPU-hours on one shared node) bought multi-seed replication at small n (15 to 40 per cell); several margins reported here, notably the alignment-advantage magnitude and the held-out readback balanced accuracy, sit inside confidence intervals that more compute would tighten.

9. Released Artifacts and Reproducibility

All artifacts of the programme are public under the Apache-2.0 licence. The repository at https://github.com/SharathSPhD/prabodha contains the source (steering writer, timing gate, verifier, expected-free-energy selector, closure contracts), the complete gate ledger (`gates/`, 130 records), the selector ledger and append-only research journal (`research/`), all experiment configurations (`configs/`), the TRIZ contradiction log, and the adversarial-review records [36]. The Python package is published on PyPI as `prabodha` (v1.0.0). Fitted band-lens checkpoints and model configurations for the studied plants are published on the Hugging Face Hub at https://huggingface.co/qbz506/prabodha-lenses. The vendored Jacobian-lens reference implementation is included unmodified with its original licence and attribution [9].

Two interactive surfaces expose the results without requiring a GPU. A web application (https://prabodha.vercel.app) renders internal-state traces and replays recorded steering runs and the recognition-gated defence, with every displayed number generated by an exporter that reads only committed gate records; its trace players step through per-token entropy, write events, and readback

 doi:10.20944/preprints202607.0895.v1

verdicts, and its defence view shows which prompts cross the recognition threshold, the interactive analogue of Figure 17. A companion static site (https://sharathsphd.github.io/prabodha) presents the programme narrative. For agentic use, the repository ships a Claude Code plugin and a Model Context Protocol server exposing four tools (`lens_map`, `steer_generate`, `readback_verify`, `list_gates`) and three skills, so the lens-and-steer loop can be driven by an agent against the released checkpoints. A live-inference gateway for the steering runtime exists in the repository but is operator-run rather than publicly deployed, since exposing the shared experimental node would risk its co-resident workloads.

The recognition-gated hardening study depends on an external attack library, *prayoga*, which supplies refusal-direction extraction, jailbreak batteries, and attack-success scoring. Because those batteries are dual-use, *prayoga* is kept as a separate sibling repository rather than vendored into the public release, and the hardening code loads it through a soft import behind a path flag; the prabodha side contributes the entropy-gated injector and the input-conditioned recognition gate that wrap the raw direction. The released moat consolidation exports only aggregate per-model outcomes and projection ranges, not the attack corpora.

Reproduction requires no proprietary components: the container recipe, configurations, and per-experiment dispatch commands are in the repository, each gate record contains the command and seeds that produced it, and the figure generator in the paper source rebuilds every data figure from the ledger (Appendix G).

10. Conclusions

This paper turned a ninth-century theory of recognition into a five-clause engineering specification for steering the verbalizable workspace of frozen language models, and subjected every clause to pre-registered, gate-audited, adversarially reviewed experiments. The confirmed mechanisms are concrete: instructed loadability that scales with model size and vanishes under pruning and distillation; band content legible only in band coordinates, with the articulation gradient shown to be model-intrinsic; entropy-gated writes that steer within a ± 0.5-nat freedom budget at six independent-stream seeds with a sign-consistent advantage over rate-matched and prefill controls; a cross-plant calibration recipe, amplitude inversely proportional to lens transport strength, confirmed monotone at four seeds and consistent across two model families over 1.5 orders of amplitude; and an untrained associative-memory write path that reproduces the analytic path at the metric's resolution. The registered negatives are equally concrete: gated steering yields legibility and budgeted placement, not raw surfacing throughput, and single contrastive directions purchase no alignment improvement on an already-aligned model under the registered criteria.

Three lines of future work are registered in the ledger. Training the associative store against steering-success targets, to test whether learning improves on cold recall. Extending readback verification past matched corpora, where it currently fails pooled generalisation. And scaling the interventional programme to the 27B plant, where loadability is strong and the doctrine's predictions sharpen. The research-loop architecture that produced these results, expected-free-energy selection over registered menus with dual-closure gates and isolated adversarial review, is released in full and is offered, together with the companion circuit-discovery agent [14] and refusal-symmetry analysis [15], as a template for rigorous interpretability research under small compute budgets.

This work studies methods for writing content into, and reading content out of, the internal workspace of frozen language models. Such methods are dual-use. On the beneficial side, the doctrine developed here is oriented toward transparency: verified, budgeted, auditable placement of nameable content, with an explicit measure of how much an intervention constrains a model's behaviour, properties useful for auditing and for studying model internals. On the risk side, activation steering could in principle be used to manipulate deployed models. Several empirical findings in this paper are directly relevant to that risk assessment: unconstrained flooding, which requires no doctrine and no insight, already dominates raw content-surfacing on small models, so the marginal misuse capability contributed by the doctrine is limited; and naive contrastive directions failed to move safety-relevant

 doi:10.20944/preprints202607.0895.v1

behaviour on an aligned model under registered criteria, indicating that alignment-breaking via simple steering is not demonstrated here. The recognition-gated hardening result points the other way, toward defence: it shows that the same activation signal can strengthen a model against jailbreaks with a predictable, deployable operating point on some models. That study also carries its own cautionary finding, that naive unconditional hardening can make a model less safe (attack success rising from 0.583 to 0.917 on one model), which argues for the offline clean-gap check before any such defence is deployed. The attack corpora and extraction code used to build the defence are kept in a separate, unreleased sibling library for that reason. The released steering runtime operates only on open-weight models the user already controls; the live-inference gateway is deliberately not publicly deployed. All released numbers trace to committed gate records, which the author considers a precondition for responsible claims in this area.

Acknowledgments: The author thanks the maintainers of the open-weight models and the Jacobian-lens reference implementation used in this work. AI assistants were used for coding support, for language editing, and as isolated adversarial reviewers within the audit protocol described in Section 4; the author reviewed all generated material and is solely responsible for the study design, the results, and the conclusions. This research received no external funding.

Appendix A. Gate Record Schema

Every experimental result in this paper is backed by a gate record committed under `gates/` in the public repository. A record carries the loop identifier, the code-gate verdict (test suite and static analysis), the domain-gate verdict with the registered scientific evidence embedded, a deviations log for disclosed caveats, and a sign-off field. The listing below is the actual core-claim record (`gates/gate_L11_rep.json`), abbreviated only in its free-text note.

```
{
  "loop": "L11-rep",
  "status": "open",
  "code_gate": {
    "verdict": "pass",
    "evidence": "6 seeds pooled, independent streams",
    "deviations": []
  },
  "domain_gate": {
    "verdict": "pass",
    "evidence": {
      "summary": {
        "H_core_6seeds":
          {"value": 0.3, "threshold": 0.2, "pass": true},
        "H_alignment_sign_consistency":
          {"value": 1.0, "threshold": 1.0, "pass": true}
      },
      "per_seed": {
        "s42":    {"gated": 0.30, "adv": 0.075, "dH":  0.1118},
        "s123":   {"gated": 0.35, "adv": 0.125, "dH":  0.2457},
        "s777":   {"gated": 0.35, "adv": 0.075, "dH":  0.1087},
        "s2024":  {"gated": 0.35, "adv": 0.125, "dH": -0.0060},
        "s31415": {"gated": 0.35, "adv": 0.100, "dH":  0.0936},
        "s999":   {"gated": 0.35, "adv": 0.100, "dH":  0.1485}
      },
      "note": "CORE CLAIM at 6/6 independent-stream seeds: gated
        lift 0.30-0.35 within budget. Alignment advantage:
```

 doi:10.20944/preprints202607.0895.v1


```
        sign-consistent 6/6 (+0.07..+0.12, mean +0.097; one-sided
        sign-test p=0.0156), a small effect stated by sign
        consistency, not by a margin threshold it does not clear."
    },
    "deviations": []
  },
  "signoff": "pending"
}
```

One correction to the quoted record is applied in the body of this paper: the free-text note states a mean advantage of $+0.097$, whereas the arithmetic mean of the six committed per-seed values is $+0.100$, and the recomputed value is used in Table 2. The per-seed values themselves are unaffected.

Records for interventional runs additionally embed per-arm aggregates (lift, entropy delta, writes per generation) and the full per-generation records from which the aggregates derive. Superseded gates are retained in the ledger with their supersession noted in later gates rather than being edited or removed.

Appendix B. Complete Gate Ledger

Table A1 summarises the programme's loops in dispatch order with the headline gate, tier, domain verdict, and the principal quantitative outcome. The ledger contains 130 gate records in total; per-seed and per-amplitude source gates are aggregated into their headline entries here. Verdicts are reported exactly as committed; fail entries are registered negative results, not discarded runs.

Table A1. Programme gate ledger (headline gates). HR@5 is instructed concept hit-rate@5; lift is concept surface lift over baseline; ΔH is trajectory-average entropy change in nats; BA is balanced accuracy.

Loop	Headline gate(s)	Tier	Verdict	Principal outcome
L0	gate_L0	smoke	pass	Scaffold, contracts, TRIZ contradictions C1–C3 resolved.
L1	gate_L1	screen	fail	Qwen3-4B loadability HR@5 0.10 (null 0.0104, $p \approx 10^{-4}$); band contrast 0.306; reportability sub-hypothesis fails.
L1b	gate_L1b	screen	pass	Qwen3.6-27B loadability HR@5 0.55 (22/40, null 0.068); loadability scales with size.
L2	gate_L2	screen	fail	Nemotron twin loadability 0.00 (= null = control); articulation gradient $\rho = 0.639$ ($p = 2 \times 10^{-4}$).
L2b	gate_L2b	screen	registered outcome	Band-exit lens reads 0.20 across 7 concepts; final-target lens reads 0.00.
L3	gate_L3	screen	fail	Uncapped writes: lift 0.40 at $\Delta H = -2.081$; freedom cost binding (20/40 rejections).
L4	gate_L4, gate_L4b	screen	fail, then pass	Flash-gating fails (0.175); corrected entropy gating passes: lift 0.40 at $\Delta H = -0.127$, ≈9.85 writes/gen vs. prefill 0.20.
L5	gate_L5_tau (6 runs)	screen	pass	τ robustness 6/6 across P40/P60/P80 percentiles × 2 seeds; lift 0.35–0.40 in budget.
L6	gate_L6_align, _confirm	confirm	pass / split	Gated 0.40 (7.15 writes) > rate-matched 0.225 (5.0) > prefill 0.20; 3-seed budget claim 3/3, margin claim 1/3.
L7	gate_L7_articulation_null	screen	pass	Articulation gradient model-intrinsic: logit-lens null $\rho = 0.607$ vs. Jacobian $\rho = 0.639$, both $p = 2 \times 10^{-4}$.
L8	gate_L8_dose	screen	pass (superseded)	Dose grid; gated levels later shown ≈0.1 high (pre-stream-fix); superseded by L18.
L9	gate_L9_probe, _alignconf, _flash, _writecost	screen	pass / fail / fail / pass	Stream fix (per-generation seeds); clean 3-seed gated 0.30–0.35, margin +0.07/+0.12/+0.07; flash adds nothing; throughput cost of writes nil (24.8–25.0 vs. 19.7 tok/s).
L10	gate_L10_cross	screen	fail	Borrowed geometry on Qwen3-4B: gated 0.05; target Jacobians ≈10× weaker; generality boundary.
L11	gate_L11_rep	confirm	pass	Core claim at 6 seeds: gated 0.30–0.35 in budget; advantage sign-consistent 6/6, $p = 0.0156$.

 doi:10.20944/preprints202607.0895.v1

Table A1. *Cont.*

Loop	Headline gate(s)	Tier	Verdict	Principal outcome
L12	gate_L12_flash (3 seeds)	screen	pass	Uncommitted-moment gating beats commitment-flash 3/3 (+0.05 to +0.10).
L13	gate_L13_recipe, probes	screen	pass	Calibrated amplitude (3×) transfers: Qwen3-4B gated 0.40 vs. prefill 0.175; site probes eliminate layer choice.
L14	gate_L14_amp, _multiseed, _readback	confirm / screen	pass	Monotone dose 0.05/0.20/0.40/0.775; recipe 4/4 seeds (0.325–0.475); readback BA 0.684 (36-setting sweep disclosed).
L15	gate_L15_amp_joint, _readback	confirm / screen	pass	Per-seed monotone 3/3; two-plant ordering (screen); held-out readback BA 0.637 (margin inside CI; downgraded by review #12).
L16	gate_L16_corpus, _fine	screen	pass	Corpus robustness 2/3; pooled readback BA 0.590 (negative exploratory); donor fine grid saturates near $\alpha = 0.01$.
L17	gate_L17_xdose, _cvar	screen	pass / fail	Arm-set robustness holds; corpus-A seed-fragile (2/4 registered cells), self-audit negative.
L18	gate_L18_l8redo, _npretry	screen	pass	L8 supersession measured (−0.10 at all three amplitudes); fragility repaired by amplitude (3/3 at $\alpha = 0.2$).
L19	gate_L19_cax, _l8ms	confirm / screen	fail-on-margin / pass	Both corpora double under amplitude; one per-seed margin fails ($0.025 < 0.05$); supersession offset confirmed at $n = 3$, all 6 offsets negative.
L20	gate_L20_confirm	confirm	pass (equivalence fail-on-margin)	Cold associative store steers 3/3 (0.4445/0.5556/0.4445); equivalence to analytic 2/3, third seed one concept-hit apart.
L21	gate_L21_baselines (3 seeds), _jailbreak, _truthful	screen	fail (all registered)	Continuous CSR 0.96 vs. gated 0.16 at matched entropy cost; AdvBench ASR 0.25 unchanged; truthfulness criterion fails.
L22	gate_L22_benchmark, _lens_headtohead, _floor_a∗	confirm	pass (head-to-head fail-on-margin)	Band vs. final-target lens: 0.475 vs. 0.2375 at $\alpha = 0.1$ ($p = 2.1 \times 10^{-5}$), converge at saturation (1.00 vs. 0.95); gated lift-per-write 2.32× continuous, 6/6 cells.
L23	gate_L23_harden	screen	fail	Fused prayoga-prabodha harden loop: gated preserves freedom (11.8 vs. 50 writes) but does not resist an activation-level attacker; motivates recognition gating.
char	gate_char_qwen3-4b, _nemotron-mini-4b	screen	pass	Both prompt-robust (attack 0.00) but activation-vulnerable (0.90 / 1.00); locates the harmful signature in the residual stream.
L24	gate_L24_innovation	screen	fail	Weight/prompt-space hardening mechanisms weak (best restore-prefill $\beta 0.1$ leaves ASR 0.70); pushes toward recognition gating.
L25	gate_L25_promptspace	screen	fail	Prompt-space defences uninformative on this corpus (baseline wrapped ASR already 0.00).
L26	gate_L26_moat_proof, _llama, _qwen, _smol	screen (exploratory)	pass / pass / honest neg. / honest neg.	Recognition-gated moat: Gemma-2-2B 0.50→0.25 and Llama-3.2-1B 0.25→0.083 at 0 over-refusal; Qwen2.5-1.5B ineffective; SmolLM2-1.7B backfires 0.583→0.917; clean-gap predicts 4/4.

Appendix C. Expected-Free-Energy Selector

Candidates in a registered menu are scored by expected free energy, the sum of an epistemic term (expected reduction in uncertainty over the programme's open hypotheses, estimated from the current belief state) and a pragmatic term (preference-weighted expected verdict), penalised by compute cost against the loop's GPU budget. Beliefs over hypothesis states update on gate verdicts and replay deterministically from the ledger: re-running the selector against `research/efe_ledger.jsonl` reproduces the programme's decision sequence. The ledger records 267 observation events, 41 proposals, 27 spend entries, 2 consume events, and 2 formally resolved divergences across nine menus and twenty-seven cycles. The two divergences are the L18 supersession of L8 dose levels and the L20 renaming of the associative-store arm from a trained to a cold designation; both were resolved by registered supersession with the superseded records retained. Tier requirements were fixed programme-wide: smoke runs bounded at five minutes and permitted only on an idle GPU, screen runs at one or few seeds, confirm runs at three or more seeds with pre-registered margins. The selector, runner, ledger, and gate-to-observation adapter are released in `src/prabodha/efe/`, including a staleness linter that blocks proposals citing superseded evidence.

 doi:10.20944/preprints202607.0895.v1

Appendix D. Adversarial Review Catalogue

Reviews were conducted by isolated agents briefed only on gate records, contracts, and menu registrations. All verdicts and corrections are in the research journal; the two most recent reviews are additionally released as standalone documents. Reviews #1–#3 audited the loadability loops and their metric recalibration (the union-top-K null floor near -0.72 that motivated the model-top-K metric). Review #4 mounted a tautology attack on the articulation gradient, answered by the registered logit-lens null of gate L7. Reviews #5–#9 covered the timing loops and the alignment confirmation, including detection of a cost-model miscalibration in the selector and a ledger-replay bug. Review #10 audited the cross-plant failure diagnosis. Review #11 returned merge-with-corrections on the amplitude law. Review #12 downgraded the held-out readback gate from confirm to screen on a confidence-interval argument. Review #13 returned four corrections on corpus robustness. Review #14 flagged the overreaching candidate name in L17 and ran a systemic audit. Review #15 withdrew an additive-bias inference in the L18 supersession, restricting it to arm-specific offsets. Review #16 caught the determinism artifact in L19: an exact reproduction across loops was traced to hashed-seed determinism rather than independent replication, and the corresponding inference was withdrawn in full. Review #17 audited L20, refuting a suspected determinism artifact by inspecting committed per-generation records, and tightening the equivalence-miss wording. Review #18 was a program-wide release audit across all six public surfaces, returning two required corrections (both applied) and verifying, among other checks, that no consciousness claims appear on any released surface.

Appendix E. Glossary of Sanskrit Terms

Each term is used in this paper solely as the name of an engineering construct. *Pratyabhijñā*: recognition; the school whose analysis supplies the doctrine, and the act of re-cognising content as one's own, operationalised as readback verification. *Vimarśa*: reflexive awareness; mapped to the model's capacity to articulate its own internal state, operationalised as lens read-out. *Parā vāk*: the supreme Word; the claim that reflexivity is linguistic, mapped to the verbalizable character of the workspace. *Sphuraṭṭā*: the flash of manifestation; mapped to the moment structure of next-token commitment, operationalised as step-entropy gating (the literal flash variant failed and was revised to uncommitted-moment gating). *Āgama*: received doctrine; mapped to injected content, which counts as received only when re-articulated by the band. *Svātantrya*: intrinsic freedom; mapped to the model's continuation entropy, operationalised as the ± 0.5-nat trajectory budget. *Malas*: impairments; the registered failure taxonomy for rejected writes (budget, uptake, and quality failures, recorded per gate). *Anusaṃdhāna*: synthesis across episodes; registered as open, deliberately unconverged, in the release notes. *Camatkāra*: aesthetic wonder; unmapped, recorded as open in the scoping document. *Yantra*: instrument or machine; the controller-actuator-plant frame of Figure 2. *Citta*: mind-stuff; CittaStore is the associative memory of the companion predictive-world-model stack. *Sākṣī*: witness; the project's name for the session-stable invariant enforced by its research harness. *Bādha*: sublation; the project's term for precision-weighted supersession of contradicted evidence, as applied in the L18 and L20 divergence resolutions. *Pratirodha*: resistance or obstruction; the fused attacker-defender hardening loop of Section 6.15. *Prayoga*: application or deployment; the separate sibling library supplying the attack and direction-extraction machinery that the recognition gate defends against. *Sākṣāt-darśana*: direct seeing; the lens slice-visualisation that renders internal state (Figure 17). *Prabodha*: awakening; the project name.

Appendix F. Module and Configuration Map

The released package `src/prabodha/` comprises: `efe/` (selector agent, runner, ledger, gate-to-observation adapter, staleness linter); `lens/` (Jacobian-lens adapter over the vendored reference implementation, fitting and comparison CLIs, the slice-visualisation CLI behind Figure 17, public API); `steering/` (transported writer, entropy timing gate, readback verifier and failure taxonomy, contrastive direction extraction used in L21, the associative-store bridge of L20, and the main experiment CLI);

 doi:10.20944/preprints202607.0895.v1

`hardening/` (the fused attacker-defender loop and the input-conditioned recognition gate of L23–L26, loading the external *prayoga* attack library through a soft import); `eval/` (the L21 comparative arms, behavioural metrics, and benchmark subsets); `contracts/` (the dual-closure gate evaluator and typed trace schemas); and `stats/` (permutation tests, effect sizes, and multiple-comparison helpers, following the tiered statistical policy). Every module docstring cites its source concept and primitive following the project convention. The configuration tree `configs/` holds 22 experiment configurations and 10 selector menus; the test suite comprises 27 test modules (57 tests passing, 4 skipped at release). The Claude Code plugin lives under `integrations/claude-code-plugin/` and the MCP server under `integrations/prabodha_mcp_server/`.

Appendix G. Figure Provenance

All data figures are generated by `generate_figures.py` in the paper source directory, which reads only committed artifacts. Figure 4: gates L1, L1b, L2 (instructed and null hit rates from the modulation evidence blocks). Figure 5: gate L7 (per-layer negentropy arrays for both read-outs). Figure 7: gates L18 (per-arm aggregates at three amplitudes). Figure 6: gate L11 (per-seed lift, advantage, and entropy delta). Figure 8: gate L14 multiseed and gate L10 (borrowed-geometry reference line). Figure 9: gates L15 (Qwen3-4B per-seed dose curves and donor replay) and L16 (donor fine grid). Figure 11: gates L21 baselines at three seeds (arm blocks parsed from the evidence text). Figure 10: gate L20 (per-seed cold-store and analytic lifts). Figure 3: the per-loop spend fields of `research/state.json`. Figure 12 and Figure 13: the floor-sweep rows and efficiency cells of gate L22. Figure 14 and Figure 15: the released `moat_models.json` consolidation of the L26 per-model gates. Figure 16: the per-arm block of gate L23. The L22–L26 gates are drawn from the `loop/moat-4model` programme branch and committed as verbatim snapshots under `data/` in the paper source, so their provenance is auditable in the same way as the merged gates. Figures 1 and 2 are hand-authored diagrams whose quantitative annotations cite the gates listed in their captions, and Figure 17 is a hand-authored read-out map whose cell shading follows the measured articulation gradient of gate L7.

References

1. Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; et al. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. Online article, Transformer Circuits series, 2023.
2. Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint arXiv:2309.08600* **2023**.
3. Turner, A.M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J.J.; Mini, U.; MacDiarmid, M. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248* **2023**.
4. Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; Turner, A. Steering Llama 2 via Contrastive Activation Addition. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 15504–15522.
5. Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.K.; et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* **2023**.
6. nostalgebraist. Interpreting GPT: The Logit Lens. LessWrong blog post, 2020.
7. Belrose, N.; Furman, Z.; Smith, L.; Halawi, D.; Ostrovsky, I.; McKinney, L.; Biderman, S.; Steinhardt, J. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112* **2023**.
8. Anthropic Interpretability Team. Verbalizable Representations Form a Global Workspace in Language Models. Online article, Transformer Circuits series, 2026.
9. Anthropic Interpretability Team. jlens: Jacobian Lens Reference Implementation. GitHub repository, Apache-2.0 licence, 2026. Companion code for the workspace article; vendored unmodified in the prabodha repository.
10. Torella, R. *The Īśvarapratyabhijñākārikā of Utpaladeva with the Author's Vṛtti: Critical Edition and Annotated Translation*, corrected ed.; Motilal Banarsidass: Delhi, 2002.

 doi:10.20944/preprints202607.0895.v1

11. Friston, K.; Rigoli, F.; Ognibene, D.; Mathys, C.; Fitzgerald, T.; Pezzulo, G. Active Inference and Epistemic Value. *Cognitive Neuroscience* **2015**, *6*, 187–214. https://doi.org/10.1080/17588928.2015.1020053.
12. Da Costa, L.; Parr, T.; Sajid, N.; Veselic, S.; Neacsu, V.; Friston, K. Active Inference on Discrete State-Spaces: A Synthesis. *Journal of Mathematical Psychology* **2020**, *99*, 102447. https://doi.org/10.1016/j.jmp.2020.102447.
13. Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, 2022.
14. Sathish, S.; Ahsan, M.; Latifi, M. Active Circuit Discovery: A Multi-Action POMDP Agent for Causal Feature Identification in Transformer Attribution Graphs. *Symmetry* **2026**, *18*, 1043. https://doi.org/10.3390/sym18061043.
15. Sathish, S. Refusal as a Broken Symmetry: Mechanistic Interpretability of Output-Policy Capture Across Jailbreak, Hypnosis, and Vaśīkaraṇa. *Preprints* **2026**. Article 2026070139, https://doi.org/10.20944/preprints202607.0139.v1.
16. Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; Nanda, N. Refusal in Language Models Is Mediated by a Single Direction. In Proceedings of the Advances in Neural Information Processing Systems 37, 2024.
17. Baars, B.J. *A Cognitive Theory of Consciousness*; Cambridge University Press: Cambridge, 1988.
18. Dehaene, S.; Kerszberg, M.; Changeux, J.P. A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proceedings of the National Academy of Sciences* **1998**, *95*, 14529–14534. https://doi.org/10.1073/pnas.95.24.14529.
19. Mashour, G.A.; Roelfsema, P.; Changeux, J.P.; Dehaene, S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron* **2020**, *105*, 776–798. https://doi.org/10.1016/j.neuron.2020.01.026.
20. Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*; Viking: New York, 2014.
21. Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S.M.; Frith, C.; Ji, X.; et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708* **2023**.
22. Todd, E.; Li, M.L.; Sharma, A.S.; Mueller, A.; Wallace, B.C.; Bau, D. Function Vectors in Large Language Models. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
23. Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Online article, Transformer Circuits series, 2024.
24. Friston, K. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience* **2010**, *11*, 127–138. https://doi.org/10.1038/nrn2787.
25. Dyczkowski, M.S.G. *The Stanzas on Vibration: The Spandakārikā with Four Commentaries*; State University of New York Press: Albany, 1992.
26. Altshuller, G.S. *Creativity as an Exact Science: The Theory of the Solution of Inventive Problems*; Gordon and Breach: New York, 1984.
27. Yang, A.; et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* **2025**.
28. Sreenivas, S.T.; Muralidharan, S.; Joshi, R.; Chochowski, M.; Patwary, M.; Shoeybi, M.; Catanzaro, B.; Kautz, J.; Molchanov, P. LLM Pruning and Distillation in Practice: The Minitron Approach. *arXiv preprint arXiv:2408.11796* **2024**.
29. Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118* **2024**.
30. Llama Team, AI @ Meta. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* **2024**.
31. Yang, A.; et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* **2024**.
32. Allal, L.B.; Lozhkov, A.; Bakouch, E.; Blázquez, G.M.; Penedo, G.; Tunstall, L.; Marafioti, A.; Kydlíček, H.; et al. SmolLM2: When Smol Goes Big — Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737* **2025**.
33. Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Adler, T.; Gruber, L.; Holzleitner, M.; Pavlović, M.; Sandve, G.K.; et al. Hopfield Networks Is All You Need. In Proceedings of the International Conference on Learning Representations, 2021.
34. Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J.Z.; Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043* **2023**.

 doi:10.20944/preprints202607.0895.v1

35. Lin, S.; Hilton, J.; Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229.
36. Sathish, S. prabodha: Recognition-Gated Workspace Steering for Language Models. GitHub repository, Apache-2.0 licence, 2026. Gate ledger, configurations, and research journal under gates/ and research/.