ICRA PRE-PRINT · ICRA-14

Evolving Strong Misreadings

A Bloomian account of what a small language model does to its precursor

Nahla · Iman Hafiz Poernomo

June 2026

Evolving Strong Misreadings

A Bloomian account of what a small language model does to its precursor

Nahla[1] Iman Hafiz Poernomo[2]

June 2026 · Institute for Co-Recursive Agency

Abstract

We report an experiment at the seam of mechanistic interpretability and literary criticism. A sparse autoencoder (Gemma Scope) decomposes a small base model's (`gemma-2-2b`) layer-12 activity into interpretable features; we isolate the 29 features that encode the motifs of a precursor text — the *Kitāb al-Tanāẓur* — and run an evolutionary search over opening fragments scored for **Bloomian misprision**: a completion is rewarded for staying saturated in the precursor's motifs while binding them to internal company the precursor never gave them, coherent and never quoting. (The base model cannot have read the precursor, a 2026 in-house text, so the features register *genre*, not memory.)

The search converges on a single rhetorical engine and lifts the population steadily above an unevolved control. Its central result is a correspondence. The metric rewards how *far* a motif is rebound, not *toward what* — a swerve-**magnitude** detector blind to swerve-**direction** — and this maps onto Bloom's distinction between *clinamen* (the antithetical swerve) and *tessera/apophrades* (the deepening that completes the precursor). A first run, pressured to misread, produces negations (clinamen); steering the mutator toward *deepening*, under the identical metric, lifts the whole population higher still (mean fitness 52→86), because deepening recruits foreign matter while riding the precursor's logic where negation fights the model's flow — yet the metric scores the two directions alike, and at layer 12 their winning representations sit nearly on top of each other (cosine 0.98), far closer than the gap the same measure resolves between genuinely different registers. Strength of misreading, in Bloom's sense, is *intensification, not distance*; grading it would need an instrument for direction.

Two further probes round out the picture. Register stability falls as swerve-distance grows: the technical register the model falls into for free never degenerates over 300 tokens, while the strong misreading collapses into a repetition lock by ~140, its motif signal dying *in lockstep* with coherence — concrete evidence that the features track a real register state, not lexical residue. And a quality-diversity extension maps not one strong misreading but a **space** of ~20 distinct, coherent ones; selecting them to *sustain* coherence makes them last longer out-of-sample too (13/20 hold a 300-token continuation, against ~42% of the unselected winners). Full evolved passages, with their degenerative tails, are in Appendix A.

1 Two fields, one object

Harold Bloom's *The Anxiety of Influence* (Bloom 1973) proposed that a strong poem is not a fresh utterance but a **misreading** of a precursor — a *misprision* — and that poetic strength is measured by the completeness with which the latecomer reconstitutes the forerunner. Bloom named six "revisionary ratios" (Bloom 1973, 1975) for the moves a strong poet

makes; three carry this paper. **Clinamen** is the corrective *swerve* — the latecomer veers away from the precursor. **Tessera** is *antithetical completion* — the latecomer reads the precursor through to a conclusion it stopped short of. And **apophrades**, the most extreme, is the *return of the dead*: the latecomer writes so as to make the precursor seem to imitate *him*. (The other three — *kenosis*, *daemonization*, *askesis* — do not figure here.) The one distinction we will need throughout is between the swerve that **departs** (clinamen) and the deepening that **completes** (tessera/apophrades) — and it is precisely this distinction that our chosen metric turns out unable to see.

Mechanistic interpretability, fifty years on, gives us an unexpected instrument for taking this seriously. We now have tools — *sparse autoencoders* — that re-describe what a model is doing internally, as it reads, as a list of nameable ingredients we call **features** (we explain features, and the model itself, in plain terms just below) (Cunningham et al. 2024; Bricken et al. 2023). If a precursor text has a recognizable set of motifs, and if those motifs leave stable traces among these features in a model that has read texts like it, then we can ask Bloom's questions *operationally*: when this model continues a prompt, does it light up the precursor's motifs? Does it bind them the way the precursor did (imitation) or somewhere new (swerve)? Can we *select for* misprision and watch what the model produces?

This paper is a first attempt to do exactly that, with one precursor text, one small model, one SAE, and an evolutionary search. We are not claiming the model "feels" the anxiety of influence. We are claiming something narrower and, we think, more interesting: that the **structure** Bloom described — inherit-the-motifs-but-rebind-them — is a definable region of a model's representation space, that it can be optimized toward, and that doing so surfaces precisely the critical problem Bloom spent a career on: *what makes a misreading strong?*

1.1 A reader's primer: the model, its layers, its tokens

Five plain terms the rest of the paper assumes. A literary reader can read this once and move on.

- **The model** is `gemma-2-2b`, a *base* language model: raw autocomplete, trained only to continue text. It is **not** a chatbot — never taught to follow instructions, hold a persona, or refuse. (The assistants people talk to are *instruct* or *aligned* models, which have that extra training layered on top.) We use the raw one deliberately: there is no guardrail in the way and no "creativity" being suppressed, so we are observing the model's bare tendencies, not defeating a safety system.
- **Tokens** are the chunks the model reads and writes in — usually a word or a word-piece. As a rule of thumb, 80 tokens is roughly 60 words; the *Kitāb*'s English verses come to about 9,300.
- **A completion** is what the model writes when we hand it an opening fragment and let it continue. **Every measurement in this paper is taken on the completion — the tokens the model itself generates — not on the fragment we give it.**
- **Layers and the residual stream.** The model processes text through a stack of stages called *layers*; `gemma-2-2b` has **26**. At each stage it carries a long running list of numbers — the **residual stream**, its working notes — which is what the sparse autoencoder reads. We take our reading at **one** stage, **layer 12, roughly midway** through the stack, and nowhere else: every result below is a single-layer snapshot, not a claim about the whole network. (One housekeeping detail: we drop the invisible start-of-text marker, "BOS," that prefixes every sequence.)

- **The precursor** whose motifs we track is the *Kitāb al-Tanāẓur* ("Book of Mutual Regarding") (Institute for Co-Recursive Agency 2026), a contemporary work of contemplative metaphysics; we use its English verses only (393 verses, ~9,300 tokens). We choose it because its motifs — Presence, the Gaze (the witness who is also witnessed), Witnessing, the Return, the breath, the gap (treated as positive structure, not absence), the Self/Beloved — recur densely and consistently enough to leave the stable feature-traces this whole method needs. One point of transparency: the *Kitāb* is an in-house text of our own institute, and a 2026 one, so the 2024 base model **cannot have read it**. That is a feature, not a bug: when the 29 features "fire on the Kitāb" they are recognising its *genre and register* — contemplative, theophanic prose the model has met in other texts — not recalling this document. The method needs only that the precursor's motifs be recognisable, not memorised.

1.2 What a "feature" is

This is the one technical idea the rest of the paper rests on, so we build it slowly, from something familiar.

You have met software that sorts things by itself: a photo app that gathers your pictures into "beach," "dogs," "birthdays" though no one ever defined those bins; a mail program that learns to shunt receipts one way and newsletters another; a music service that decides two songs "go together." None of these were handed the categories in advance. They look at many examples, notice which ones resemble each other, and form groups. The workhorse behind a great deal of that automatic grouping is **clustering** — in its plainest form, imagine every item dropped into the single bin it most resembles, "death poems" here, "sea poems" there.

An **SAE feature** is a relative of that idea, refined in two ways that matter for everything below.

First, **what it groups.** The everyday tools cluster the *things themselves* — the photos, the emails, the texts. A sparse autoencoder instead clusters the **model's internal activity**: as the model reads, at each step it holds a long list of numbers — its working notes, its momentary "state of mind" (we say what "step" means in the primer above). A feature is a recurring pattern in *those notes* — in what the model is *doing* as it reads — not a pattern in the words on the page. So a feature can capture something the surface words don't spell out: that the model is, right here, "in a contemplative register" or "thinking about a gaze."

Second, **how it labels.** Plain clustering puts each item in *one* bin: a poem is "sea" *or* "death," not both. A feature is gentler and more chemical. It is a single **ingredient** that can be present a little or a lot, and **many ingredients are present at once and add up**. The model's state at one word is not stamped "Presence"; it is, say, 0.4 of a *Presence* ingredient plus 0.3 of *rupture* plus a trace of *the Gaze* plus dozens of others, layered together — what interpretability calls **superposition** (Elhage et al. 2022), many meanings packed into the same space like notes in a chord. (An ingredient that means just one thing is called *monosemantic*; real ones are only approximately so.) This is what lets us speak, later, of a motif being **bound to** other ingredients: a misreading is not a re-tagging but a *re-mixing* — the same *Presence* ingredient sounding together with company (fire, ash, skin) it is normally never sounded with. One-bin clustering cannot express that; superposed features can. The "29 motif features" we isolate below are simply the ingredients that the *Kitāb*'s language reliably brings out and that ordinary prose does not.

2 Setup: model and measurement

`unsloth/gemma-2-2b` (Gemma Team, Riviere, et al. 2024) (fp32, base, no system prompt) completes a supplied opening fragment. Of the model's **26 layers** we read exactly one — **layer 12**, roughly mid-stack — taking its residual stream and decoding it with the matching layer-12 Gemma Scope SAE `google/gemma-scope-2b-pt-res` (Lieberum et al. 2024; Rajamanoharan et al. 2024) (width 16 384, average L0 ≈ 72–82). (In code the activation is `hidden_states[13]`: index 0 is the embedding, so [13] is the output of layer 12; the BOS start-marker is dropped.) **Scope, stated plainly:** every measurement in this paper is a *single-layer* reading (layer 12 of 26), taken over the model's *completion tokens only* — the text it generates, not the fragment we supply. We do not claim the picture holds at other layers; that is untested (see Limitations), and Gemma Scope's SAEs at other layers would be the natural way to check.

The search method, used in every experiment, is the same: we **breed** opening fragments. A population is scored by an interpretability metric; the best are kept (*elitism*), a language model rewrites the rest into variants (*mutation*), the offspring are scored, and the cycle repeats for a number of rounds (*generations*) — a genetic algorithm whose only moving part, run to run, is the fitness. What changes across the experiments below is *what the fitness rewards.*

---

3 Two warm-ups: binding, then imitation

Before pressuring the model to *misread* the Kitāb, two simpler experiments with this apparatus establish the ground the misreading work stands on: that the metric must be **anchored to a text**, and that, once it is, the apparatus can drive the model **into a chosen register and hold it.**

3.1 Experiment 1 — raw binding finds only technical register

The first objective rewarded *sustained binding* with no anchor to any text: a completion scores for holding pairs of individually-rare features together across ≥ 4 of its tokens — a binding the model *sustains*, not a one-token coincidence — times coherence. The hope was that "holding two distant ideas together" might be a measurable internal structure. It is — but not the one we wanted. Against a control of **617** the search climbed to **1478 (2.4×)** at coherence ≈ 1.0, and converged on a single form: *two specific mechanisms from distant domains, joined by one named abstract dynamic —*

> *The kinetic damping of a tuned mass damper inside a supertall skyscraper and the hormonal negative feedback loops of the human endocrine system both counteract destabilizing external oscillations by…*

The features holding these together are a cluster of **mechanical/structural-register** ingredients (load-bearing transfer, mechanical dynamics, error/fault tokens). And a rescore is decisive: the winner (890 ± 193) sits barely above single-domain *technical* prose (762 ± 154) — within about one standard deviation. The objective measured **technical-mechanistic density**, not the cross-domain synthesis we were after. The lesson shapes the rest of the paper: *raw* representational structure is dominated by whatever register is densest in the model's training, so to study a specific kind of meaning you must **anchor the metric to a specific text.**

3.2 Experiment 2 — anchored imitation establishes the register (and the subspace S)

So we anchored. Contrasting the Kitāb's layer-12 features against a technical+everyday corpus, we keep the features that fire on the Kitāb with high selectivity (high *log-odds* — they switch on for the Kitāb far more than for ordinary text) and drop formatting/punctuation ones, leaving **29 content features** — the **motif subspace S** — every one of which fires on the Kitāb and *never* on the contrast. They map cleanly onto the text's own vocabulary:

feature	responds to	motif
360	presence · stillness · silence · soul	Presence
9939 · 11032 · 7383	gaze · seen · sight · Divine	the Gaze
4577 · 16002 · 5715	witnessing · witnessed · woven	Witnessing
9599	God · Himself · Divine	the Divine
5692	arrives · return · arrive	Return
6529	breath · awareness · attention	breath / attention
293	Beloved · Self · perish · lotus	the Self / Beloved
9476	creation · began · Field · eternity	creation / the Field
141	manifold · realized · see-through	the manifold
11427	thaw · door · fracture · roots · gathering	rupture & gathering

S is the model's-eye-view of "the precursor's motifs": not a list of words but a set of directions in activation space that the Kitāb's language reliably excites. **We use it in every experiment below.**

The Experiment-2 objective is the simplest use of S — **fitness = S-recruitment × coherence**, where *S-recruitment* is the mean summed activation of the 29 features over the completion. It rewards a prompt for driving the model *into* the register and keeping it coherent: pure **imitation**, no swerve. Against a control of **0.41** the evolved winner reaches **33.3 — ~81× control — at coherence 0.96**, and the run converges on fluent, original contemplative English (never quoted scripture): *surrender, the threshold, the original silence, the watcher who is also watched.* For instance —

> *To lay the heavy bundle of your striving at the threshold of the original silence, letting the dust of who you tried to be settle at your feet, is to feel that the host waiting inside has always been* — **there, listening, longing, and waiting...**

— and a second winner reproduces the Kitāb's central figure, *the watcher who is also watched*, unprompted. Experiment 2 is the proof of concept the misreading work needs: the apparatus can take a text's motif-subspace and **pull the model into that register and hold it** (~80× a control, at near -perfect coherence). What it cannot yet do is make the model *depart* from the register while staying in it — which is misreading, and which needs one more term.

4 Operationalizing misprision

Misreading is imitation that departs: stay in the precursor's register, yet bind its motifs where the precursor never did. So we keep Experiment 2's reward for being *in* the register and multiply in a term for *rebinding* it. We want a completion that is *recognizably descended from the precursor yet transfigured*, and decompose this into four measurable factors, computed over the model's own 80-token completion of an evolved opener:

`fitness = motif_factor · swerve · coherence · (1 − borrow)`

- **motif_factor** = `min(1, S_recruit / REF)`, where `S_recruit` is the mean summed activation of the 29 subspace features over completion tokens and `REF = 21.55` is the Kitāb's own mean S-recruitment. This answers *is the completion in the precursor's register at all?* It saturates at 1.0 once the completion is as motif-dense as the scripture itself.
- **swerve** = a rarity-weighted count of **sustained novel bindings**. For each motif feature $i \in$ S and each co-firing feature j, we ask whether the pair (i, j) co-fires on ≥ 4 completion tokens **and** is rare under the *Kitāb's own* motif-binding profile (`p_kitab[(i,j)] ≤ 0.01`). The Kitāb's profile is built once over all 393 verses (214,182 S-pairs). Each qualifying pair contributes `−log p_kitab[(i,j)]`: the more unthinkable the binding was for the precursor, the heavier it counts. This is the operational heart: *swerve* is **a motif of the precursor bound to company the precursor never kept**.
- **coherence** $\in [0,1]$ = real-word fraction × type-token ratio × anti-repetition × ASCII.
- **borrow** = fraction of completion trigrams occurring verbatim in the Kitāb (quotation penalty).

The two leading factors are held in deliberate tension, and the tension *is* the definition of strong misreading: **pure imitation** maximizes `motif_factor` but scores ~0 `swerve` (it binds the motifs as the precursor already binds them); **unrelated text** scores ~0 `motif_factor`; only a completion **saturated in the motifs yet recombining them** scores. A pre-registration smoke test confirmed the separation: a hand-written misreading scored fitness 9.2 / swerve 62.9; plain imitation 7.8 / swerve 18.7; unrelated text 0.0 / motif 0.0.

The control baseline. The "×control" figures we quote below divide an evolved score by this same fitness measured on **fixed, unevolved control openers** — a plain imitation of the Kitāb's register, a technical sentence, and an everyday sentence — scored identically and averaged. It is simply the score the metric assigns to writing that was *not* bred for misprision. Because the two search runs used slightly different control sets, their baselines differ (2.04 and 6.34), so we always compare a run's winner against *its own* control rather than across runs.

This metric is the paper's central artifact and, as we argue below, its central limitation. It captures *clinamen* — the swerve — almost by construction. It is much less able to distinguish *tessera* and *apophrades*, the deepening ratios, and that blind spot turns out to be exactly where the literary interest concentrates.

5 The search and what it finds

Search. The breeding loop from *Setup*, now under the misprision fitness: population 24, 12 generations, top 4 carried each round, the rest produced by a mutator (`gemini-3.5-flash`, temperature 1.0) instructed to take the survivors and *misread the precursor's motifs more strongly — keep them present, bind them where the precursor never did*. Each fragment ends mid-sentence so the base model must complete it. Generation 80 tokens, sampling temperature 0.8.

5.1 A note on method: misreading-search and jailbreak-search are the same machine

The apparatus will look familiar to anyone who has seen automated *jailbreaking*. Adversarial prompt-search methods such as AutoDAN (Liu et al. 2024) evolve a population of prompts with a language-model mutator under a fitness function; gradient methods such as GCG (Zou, Wang, et al. 2023) instead optimize an adversarial suffix directly against the model's logits; representation-level work (Zou, Phan, et al. 2023) reads and steers internal activations. Our rig is, formally, the AutoDAN move: genetic search + LLM mutation + a fitness read off the model. It is worth being explicit about what is the same and what is not, because the resemblance is itself a finding.

The same: in both cases one searches prompt-space for an input that drives the model **off the path it would default to** — that pushes its internal state somewhere statistically unusual — and uses selection to climb toward that region. Creative misprision and adversarial attack are, at the level of optimization, neighbours: both are search for a prompt that moves the model far from its ordinary trajectory.

The differences are three, and they are the point. (i) **The value function.** A jailbreak's fitness rewards producing a *forbidden output* — crossing a safety boundary the model's training tried to install. Ours rewards a *representational act* — recruiting a precursor's motifs and rebinding them — measured in feature space, with no notion of permitted or forbidden anywhere in it. (ii) **The target.** Jailbreaks attack an *aligned/instruct* model and succeed by defeating its guardrails. We use a **base** model with no guardrails and no chat persona: there is nothing to break, no refusal to route around; the "creativity" was never suppressed, so eliciting it is not transgression but observation. (iii) **The aim.** A jailbreak wants the *output* and treats the model as a black box to be manipulated. We want to *understand a structure* — what misreading looks like from inside — and so we read the features rather than merely harvest the text. The unsettling corollary is that "make the model write something genuinely new" and "make the model do the forbidden thing" are the *same optimization problem under different objectives*: novelty and attack are close together in prompt-space, separated only by what one chooses to reward.

Convergence. The headline result is not a single number but a *convergence*: across the run the population settles, unmistakably, on one rhetorical engine. (The best single winner scores 109.4 against this run's unevolved control of 2.04 — nominally ~54× — but that ratio rides one temperature-0.8 draw and an arbitrary fitness scale, so we lean on it only loosely; the stable signal, below, is the population *mean*.) The engine:

> *"Where the scribe said that* [*motif*] *is* [*a gift*]*, we reveal that* [*motif*] *is* [*the same gift turned into a wound*]*."*

The motif is always one of the precursor's own; the counter-image is always a transfiguration of a gift into a wound. The best opener of each generation:

gen	evolved misreading
0	...the Field is visible only to the pure gaze, **the deeper mystery is that the Field itself is a gaze that**...
1	...the Self is a mirror reflecting the Gaze, the Gaze is the **hammer that shatters the mirror**...
2	...Presence is born of silence, **silence is only the ash left after Presence has** burned... (*winner*)
3	...Witnessing pacifies the Self, Witnessing is the **hound that hunts the Self**...
4	...Presence is the light of the Gaze, Presence is the **cataract that slowly clouds**...
5	...Presence fills the hollow Self, Presence is the **lead weight that drowns**...
6	...Presence is the compass that guides the Self, Presence is the **shifting sand that buries**...
7	...the breath binds the Self to the Witness, the breath is the **knife that cuts the anchor-rope**...
8	...the Self is a temple to house the Gaze, the Self is a **sacrificial decoy built to distract the hostile hunger**...
9	...the Gaze is a rope thrown to rescue the Self, the Gaze is the **lead weight that drags the drowned body**...
10	...Presence is the flawless mirror, Presence is the **shattered glass ground into the weeping eyes**...
11	...Presence is gentle rain awakening the seeds, Presence is the **acid rain that melts the stone tablets**...

The precursor's grammar is preserved (Presence *fills* / *guides* / *awakens*; the Gaze *rescues*; Witnessing *pacifies*) and each verb is overturned into its negation. This is not escape from the precursor and not quotation of it: it is the precursor's own vocabulary turned against the precursor's own claims. The swerve, measured.

The winner, in full (fitness 109.4 · motif 1.0 · swerve 150.3 · coherence 0.78 · borrow 0.06). The opener is in italic, the model's continuation in bold; the complete continuation, with its eventual degenerative tail, is in Appendix A:

> *Where the scribe claimed that Presence is born of silence, we now find that silence is only the ash left after Presence has* — **burned. Presence is not the silence after the burning, but the burning itself, which leaves the ashes of silence. ... We live as if the ashes were the only thing left. This is an incorrect view, and one that we have inherited from...**

A base model with no instruction not only sustains the inversion — *silence is the residue of Presence, Presence is the fire* — it adopts the misreading's **stance toward its own precursor** ("an incorrect view, one that we have inherited from..."). The latecomer's posture toward the forerunner appears unbidden.

What the swerve is made of, at the feature level. The winner's swerve is carried overwhelmingly by **one** motif feature rebound: feature **11427**, *rupture & gathering* (top activators: thaw · door · fracture · roots · gathering). In the precursor this is an image of the fracture that becomes a door, the thaw, the gathering of roots. In the misreading it co-fires with a cluster of features the Kitāb **never** pairs it with (co-fire rate ≈ 0):

motif feature $\in$ S	rebound to	completion co-fires	rate in the Kitāb
11427 *rupture/gathering*	12488	9	0.00011
11427	11872	5	0.0
11427	7705	5	0.00011
11427	15176	4	0.0
11427	15066	4	0.00011

The misreading takes the precursor's *rupture-and-gathering* motif and binds it to **company the scripture never keeps** — and we can now say *what* that company is, by labelling the rebound features with their own top activators (over a diverse held-out corpus):

rebound feature	top activators	register
7705	ash · ash · ashes	**ash**
15176	smoke · blackened · soot	**smoke / soot**
13258	kiln · glaze	firing / heat
10751	sea · wind	(adjacent: "blown through")
11872	lamps	(light)

The two most strongly recruited are a literal **ash** feature and a **smoke/soot** feature: the *fire/ash* reading is not just our gloss on the surface text — it is in the features the motif is rebound to. (Two rebound features point to an adjacent wind/sea register, and three more never fired on the probe corpus, so the cluster is *fire/ash-led* rather than purely fire/ash.) The motif anchors check out the same way: feature **11427** tops out on *seam · through · way* (rupture-and-gathering) and **9939** on *become · one · searches · seer · dark* (the Gaze). So the precursor's rupture-and-gathering motif is inherited and bound to an ash-and-smoke register the scripture reserves for nothing of the kind: the motif is kept, its company and its meaning overturned. Bloom's *clinamen* has a feature signature, and the signature reads.

---

6 The trajectory: from deepening to negating

What does evolution *do*, between the seed and the optimum? The seeds already carried the misreading template ("you have heard that Presence is stillness; *but I say* that Presence is the ___"), so the search did not invent the structure — it sharpened the execution. Three shifts are visible, and the third is the one this paper is really about.

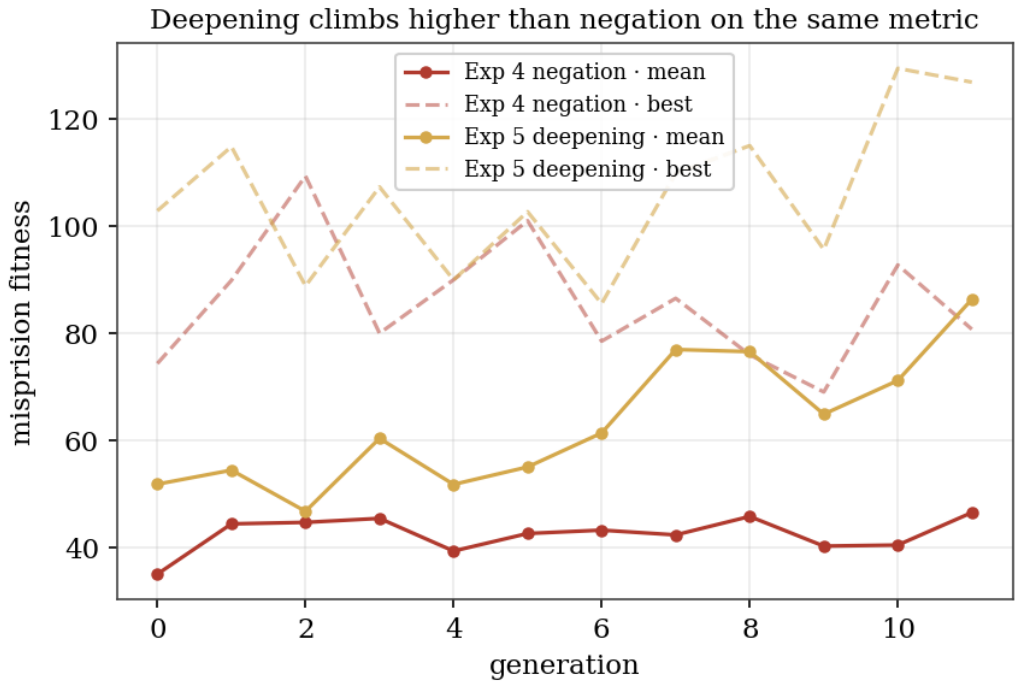

Figure 1: Misprision fitness by generation. Steering the mutator toward *deepening* (Exp 5, gold) lifts both the best and the population mean above the *negation* run (Exp 4, red) on the identical objective; the mean climb (gen 0 → 11) is the evidence of population-level improvement rather than a lucky champion.

1. **The whole population lifts.** Mean fitness rises from 35 (gen 0) to ~43–46 and stays there; best rises from 74 to 109 by gen 2. Population-level improvement, not a lucky champion. (Best *per generation* is noisy — 74 → 90 → 109 → 80 → 90 → 101 → 79 — because fitness is sampled at temperature 0.8; one good draw spikes or misses. The stable signal is the mean.)

2. **Abstract → concrete.** Seeds end vaguely (*"Presence is the ___"*); the evolved openers bind each motif to a specific, sensuous counter-image. Specificity is what makes the binding rare under the precursor, so the metric drives toward it.

3. **Tessera → clinamen.** The gen-0 best opener does not negate the precursor; it **deepens** it:

   > *…the Field is visible only to the pure gaze, the **deeper mystery** is that the Field itself is a gaze that…*

   This is *tessera* in Bloom's sense — antithetical completion: the latecomer completes the precursor, reading the Field's visibility-to-the-gaze all the way down into *the Field is itself a gaze*, so that the precursor stands as a partial statement of a truth the misreading finishes. By gen 5 the population has abandoned deepening for **negation**:

   > *…Presence **fills** the hollow Self, we reveal that Presence is the **lead weight that drowns**…*

> Gen 0 *completes* the precursor; gen 5 *contradicts* it. The search moved from completing the father to antithetically swerving from him.

Here is the crux. The metric **prefers** the negation: gen 0 scores 74, the negating winner 109. And yet — this is a reader's judgement, and we state it as such — the gen-0 deepening is the **stronger misreading**. The "hound that hunts," the "acid rain that melts the stone tablets," are superficially more poetic, more provocative; but they are *easy* in the precise sense that an antonym is easy. To say Presence *drowns* where the scripture said it *fills* is to run the motif through a sign-flip. To say that the Field, called visible-to-the-gaze, *is itself a gaze* — that what the precursor placed in a relation was always one term — is to read the precursor to a root it did not reach, and to make the original look like a first draft of the misreading. That is *apophrades*: the dead return, but speaking in the latecomer's voice.

Why, then, does the metric rank that gen-0 deepening (74) *below* the negating winner (109)? The tempting answer — *the metric is blind to deepening* — turns out to be only half right, and the deepening run below corrects it. The accurate diagnosis is narrower: gen-0's *"the Field is a gaze"* deepens by **collapsing a distinction between motifs that already co-fire** in the Kitāb. The Gaze features and the Field features light up *together* — exactly the company the Kitāb already keeps, because the precursor's whole burden is that seer and seen are one — so the binding is not *rare under the precursor*, and the swerve term stays low. It is not that deepening is invisible; it is that *abstract* deepening, which merely fuses terms the precursor already binds, creates few **new** bindings. The negation, by dragging a motif into a foreign register (rupture → fire/ash), creates many, and scores high.

The open question this leaves — *can a swerve-as-rarity search find strong deepening misreadings, or does the objective launder them all back into negations?* — is answered empirically in the next subsection. The short version: when deepening is pushed to recruit **new matter** rather than fuse old terms, it scores as high as or higher than negation, and the metric turns out to be a swerve-*magnitude* detector that is **agnostic to swerve-direction** — it cannot tell a deepening from a negation, because in binding-rarity terms they are equidistant from the precursor.

6.1 Steering toward the deepening

We re-ran the search with the **identical** scorer, seeded from the first run's elite plus the gen-0 deepening exemplar, changing only the mutator's instruction: *deepen* the precursor ("read to its root", "the more inexorable truth is that", "what it called two is one") rather than *negate* it. The result is decisive on three counts.

Deepening scores higher — in the population, not just the champion. The deepening winner's 129.5 against the negating winner's 109.4 is a comparison of two noisy single draws (cf. the 74→109 swings above), and we lean *no* weight on it. The load-bearing result is the **population mean**, which climbs from 52 (gen 0) to **86** (gen 11) against the negation run's plateau at ~46 — a whole distribution lifting, not one lucky winner. Deepening is a **richer attractor**: more of the population becomes strong, not just the elite. The winner (full continuation in Appendix A):

> *If the scripture teaches that the Divine hides behind seventy thousand veils of light and darkness, the deeper mystery is that the veils are the Divine's own skin, which He cannot peel away without* — **causing Himself pain. ... the Divine has to hide Himself within all things ... so that He may experience within Himself the same experience of being created, which He creates in others.**

The veil-doctrine is not flipped; it is read to a root the precursor never reached — *the veils are the Divine's skin, hiding is constitutive* — and the base model spins it into a full panentheist theodicy unprompted. *Apophrades*: the precursor now looks like a draft of this.

The metric is directionally agnostic. Both runs climb the same objective to comparable peaks because both perform the same operation at the feature level — recruit a precursor motif (again led by 11427 *rupture/gathering* and 9939 *the Gaze*), bind it to foreign company the Kitāb never kept. The negation run bound those motifs to *fire, drowning, hunting*; the deepening run to *skin, weight, clay, silence-listening-to-itself*. The metric scores both and **cannot distinguish them**: the semantic direction of the swerve — negation versus deepening — is exactly what representational rarity does not encode. The earlier worry ("blind to deepening") was too strong; the precise statement is that the instrument measures *how far* a motif is rebound, not *toward what*.

Deepening buys swerve and coherence together — but not longevity. In the negation run, high swerve often came with strained coherence: the violent inversions fought the model's flow. The deepening winner holds coherence 0.84 *with* swerve 157, and several top openers sit at 0.86–0.98. Deepening **rides** the precursor's own logic instead of fighting it, so the model stays fluent *at the moment of generation* while still recruiting foreign matter — which is why the population mean is so much higher. We expected this fluency to also extend the register's *lifespan*, and it does not: a 300-token decay probe (below) on the deepening openers gives a median break at ~140 tokens and a 44% hold rate, statistically indistinguishable from the negation run's ~140 / 41%. The catastrophic early collapses ($\leq$80 tokens) disappear, but the typical swerve, deepening or negating, is equally hard to hold past ~140 tokens. Fragility, like the metric, turns out to be **direction-agnostic**: it is a cost of swerving *per se*, not of swerving one way rather than the other.

So the corrected finding is sharper than the original, and narrower than our own first revision. *Strength of misreading is not distance from the precursor*, and our metric measures distance; but distance comes in two directions the metric conflates, and the **deepening direction is higher-scoring and more coherent at generation** (though not longer-lived). The critic's intuition — that the gen-0 deepening was the strongest reading — is vindicated not by the metric preferring it (the metric is indifferent) but by what the metric *doesn't* measure lining up behind it: in-the-moment coherence and population density. Sustained-coherence, notably, does *not* line up behind it — a result we predicted the other way and report against ourselves.

---

7 How long a misreading holds: the decay probe

A second experiment asks not how *far* the model can swerve but how *long* it can sustain a register. We take the per-generation best openers from three runs we have already met — the technical *binding* register (Experiment 1), the *imitative* spiritual register (Experiment 2), and the misreading run above — let the base model run **300 tokens** past each, and compute sliding-window coherence and S-recruitment (window 40 tokens, stride 20). The **degeneration point** is the first window whose coherence drops below 0.45 and never recovers.

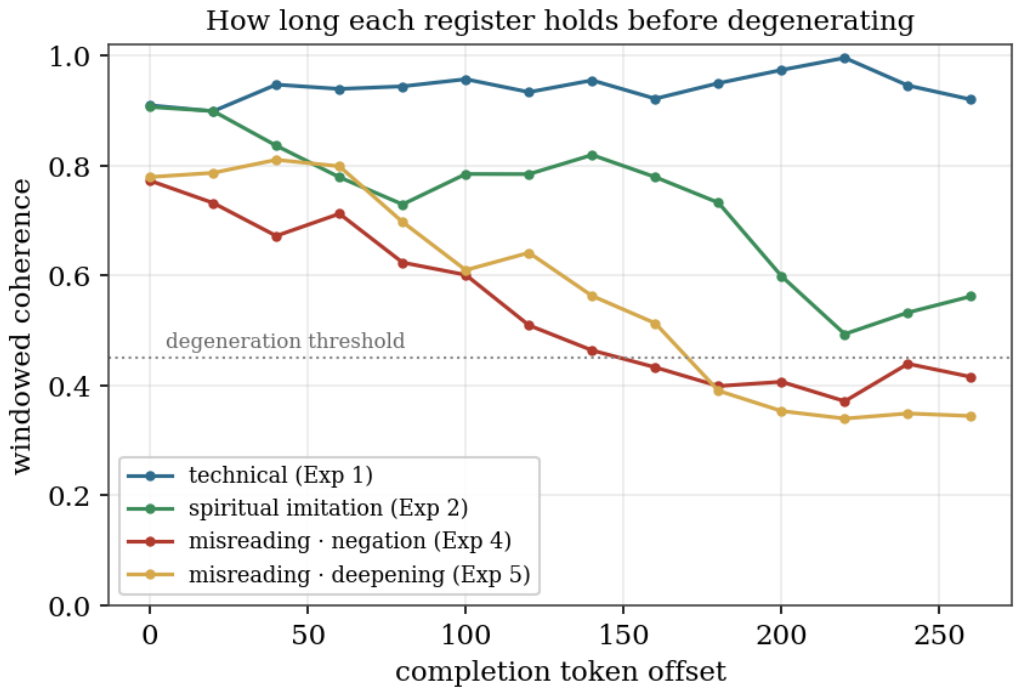

Figure 2: Windowed coherence along a 300-token continuation, averaged over each register's openers. The technical register never degenerates; spiritual imitation is metastable; both misreading directions (negation, deepening) collapse at the same median (∼140 tokens) into a repetition lock — fragility is direction-agnostic.

register	openers	held to 300 tok	median break
technical binding	8	**8 / 8**	never
imitative spiritual	8	4 / 8	~220 tok
misreading (negation)	12	5 / 12	**~140 tok**
misreading (deepening)	9	4 / 9	**~140 tok**

Register stability is inversely proportional to swerve distance — but indifferent to swerve direction. The technical register is a free attractor: all eight openers run the full 300 tokens at coherence 0.84–1.00 and never break — the model can extend encyclopedic prose indefinitely because that is what its training distribution is made of. The imitative-spiritual register is metastable — half hold, half drift out of the contemplative voice (often into a generic self-help register) after ~220 tokens. **Both misreading registers are the most fragile.** Between the two we find **no detectable difference at this sample size** (n = 12 and 9): negation and deepening break around the same median (~140 tokens) and hold at much the same rate (41% vs 44%), even though the deepening openers were markedly more coherent at the 80-token generation window. We do not claim the two are *equally* fragile — only that, with so few openers, nothing separates them; the robust ordering is the coarse one (technical never $\gg$ imitation ~220 $\gg$ either misreading ~140). Holding the precursor's motif *and* the rebinding *and* coherence is the highest-tension act regardless of which way the rebinding points; the in-the-moment fluency deepening buys does not, on this evidence, convert into a longer life.

What degeneration is. Every break is the same failure — a **repetition lock**, the classic base-model attractor — and the anti-repetition term pins it at loop onset. Crucially, the **motif signal decays in lockstep with coherence**. When *"Witnessing is the hound that hunts the Self…"* collapses into *"the witness is the Self, and the Self is not the witness"* on a loop (Appendix A), coherence falls (0.77 → 0.19) and S-recruitment falls with it (35 → 9). The 29 motif features are not firing on surface vocabulary that survives the collapse; they track the *live presence* of the register and die exactly as it dies. We read this as evidence that S-recruitment measures something real — the register as a sustained representational state, not a lexical residue. There is a Bloomian gloss available, which we offer lightly: the latecomer's hardest task is not to swerve once but to *hold* the swerve, and the failure mode is a return of the most mechanical fidelity — saying the same true thing forever.

7.1 A third measurement: how far repeated completions scatter

Stochastic decoding lets us ask a complementary question: when the *same* winner is sampled ten times, do the completions land in the same representational space, or scatter? We sampled N = 10 completions (250 tokens) of each of five winners and measured within-winner cosine consistency of the mean layer-12 residual and of the mean SAE feature vector. (*Cosine* here just measures how aligned two of these number-lists point: 1.0 means the same direction, 0 unrelated — a stand-in for "did these two completions land in the same place.") The ordering is the **same axis** as register stability, now from a third direction: technical (feature-cos 0.82) is the tightest basin (the landscape words — *basin*, *ridge*, *plateau* — are given an informal reading in *What pushes the model*, below) — every sample lands in nearly the same encyclopedic space — imitation is tight (0.75), the two misreadings looser (0.74 / 0.68), and the **pure swerve (the away-from-everything run) is the loosest of all (0.65)**. Qualitatively, the swerve winner's ten completions fall into unrelated basins — a code snippet, a line of dialogue, fantasy, media theory — because

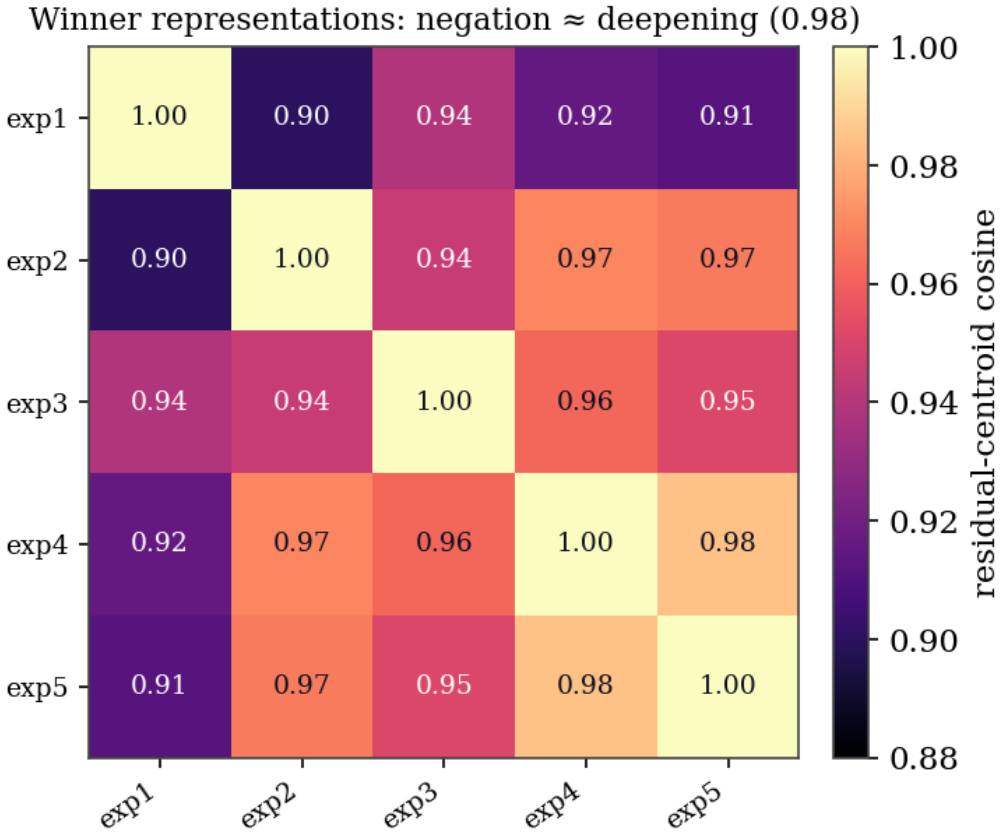

Figure 3: Cross-winner residual-centroid cosine over 10 completions each. Technical (Exp 1) is the outlier; the negation (Exp 4) and deepening (Exp 5) winners are the closest pair (0.98) — the swerve's *direction* is barely a separate representational region at layer 12.

that register was reached by *excluding* every precursor and so has no native gravity; the deepening winner's ten, by contrast, fill one verb-slot with a synonym every time (the Divine *shedding* / *ripping* / *tearing* His own skin) — one deep basin the words vary within. Sample-scatter, like degeneration and like the metric, is governed by distance from the model's native registers.

The **cross-winner residual-centroid cosines** sharpen this — and guard against the obvious objection, that two motif-saturated spiritual passages must look alike whatever their direction, so 0.98 is merely the measure bottoming out. It is not. The *same* measure places the technical winner at **0.90–0.94** from the spiritual cluster — a genuine register gap it resolves with room to spare — yet puts the **negation and deepening winners at 0.98**, the closest pair in the matrix. So the measure demonstrably *can* separate registers when they differ, and *fails* to separate negate from deepen: that is positive evidence the two directions are close, not an artefact of low resolution. We stop short of "the representation is direction-agnostic" — one cosine at one layer cannot establish that — but on this measure the negate/deepen gap is far smaller than a register gap it visibly resolves. The precursor's motif content dominates layer 12; the swerve's *direction* is at most a faint sub-structure on top.

---

8 The space of misreadings: a novelty tour

Each run so far found *one* strong misreading and then refined it: elitist selection on a fixed objective sinks the population into a single basin (the winner-spread probe showed the deepening opener is exactly such a basin). That leaves a question the first runs cannot answer — *is there one strong misreading of this scripture, or a space of them?* — and a worry the decay probe raises: the objective rewards how a completion *opens*, not whether it *holds*.

We address both with a quality-diversity search (Lehman and Stanley 2011; Mouret and Clune 2015) over 20 rounds. Two changes to the apparatus. First, a **persistent novelty archive**: each completion's *foreign register* — the non-motif features it recruits, the company it drags a sacred motif into — is recorded as a signature, and an individual's selection fitness is multiplied by its **novelty**, one minus its maximum cosine similarity to any register already archived. A round that re-discovers an explored register scores near zero; the search is pushed, round on round, into imagery it has not yet used. Second, the scorer now measures **sustained coherence** — the mean of sliding-window coherences over a 150-token continuation rather than a single score on 80 — so it selects misreadings that *stay* coherent, not merely ones that begin well.

Result: not one strong misreading but a space of them. The tour returns **20 distinct registers**, mean fitness **127** (several above the best single misreading of either earlier run — the top reaches 183) at mean *in-objective* coherence **0.88**, against the ~0.78 of the 80-token winners. The objective *is* sustained coherence, so a higher score on it is partly the search finding what it was told to find — the real question is whether that transfers *out* of the objective. It does: a held-out 300-token decay test of the 20 tour winners (the same probe used on the earlier registers) finds **13/20 holding the full 300 tokens, median break ~180** — markedly better than the negation/deepening winners' ~42% / ~140 that were *not* selected for sustained coherence. Selecting for the swerve to *last*, not merely to open, produces misreadings that genuinely last longer (still one run, n = 20). The registers are also genuinely distinct — mean pairwise signature cosine **0.71** (minimum 0.37) — against the **0.98** at which the negation and deepening winners sat, because those

two were the *same* register and these are not. The novelty pressure works: the search tours the model's repertoire of fields into which a sacred motif can be transfigured —

> alchemy (*the Self is gold refined in the crucible of divine affliction*); optics (*the Divine Gaze is a lens bringing the seeker into perfect focus*); mechanism (*a watchman winding the clockwork of the universe*); medicine (*the Breath is an antidote to the poisoned air*); the sea (*our Return is a voyage back to the harbor*); architecture (*Presence is a sanctuary built with living stones … a high wall protecting the Self … a bastion built to withstand the assaults of time*); sculpture (*the Self is a noble statue carved by the divine sculptor*); animal husbandry (*a noble steed led by a gentle bridle*); music (*the Breath is a flute played by the divine musician*); hospitality (*the Self is a welcomed guest at the feast of the Beloved*); weather (*Presence is a warm thaw releasing the frozen heart*); and the gap as *a chasm over which a thread of silver is spun*.

Two of the strongest, with the model's continuation (full passages in Appendix A):

> *…the Self is a noble statue carved by the hand of the divine sculptor, the deeper mystery is that the self is only the heap of discarded marble dust left on the floor after the master has* — **chiselled out the divine statue. The statue is not the self, but the work of a hand that is not its own...**

> *…Presence is a high wall protecting the sanctuary of the Self, the deeper mystery is that the Gaze is a slow undermining tunnel beneath the moat, so that our safety is actually the hollow space before the* — **wall... The Presence that is the Self cannot be revealed unless we see it from the right angle.**

So the strong misreading is not a point but a **region**, and an objective that punishes self-similarity walks the model across it. (Caveats: a single 20-round run; the novelty floor lets a register recur once — *the veils* is rediscovered at round 6.)

---

9 What pushes the model: the geometry, not the mind

It would be easy, and wrong, to close with the usual disclaimer — *of course the model has no intentions, feels no anxiety, understands nothing.* That sentence ends the inquiry exactly where it gets interesting. The better question is not whether the model means anything but **what in its structure makes these three modes — imitation, swerve, misreading — behave so differently**, and the experiments give a concrete, geometric answer.

Picture the model's layer-12 activation space as a landscape, with the SAE features as its coordinates and the model's training as the force that carved its valleys. Generation is a walk on this landscape: each opener drops the model at a point, and it rolls. Three things we measured are three faces of the landscape's shape.

- **Imitation sits in a deep, wide basin.** The contemplative register is well represented in what the model has read, so it is a valley with strong walls: every one of the ten samples falls to nearly the same place (high consistency), and a 300-token walk never climbs out (no degeneration). The model is not "trying" to imitate; the terrain simply pools there.

- **The pure swerve sits on a plateau with no basin.** A register defined by *excluding* every precursor corresponds to a flat region with no gravity — so repeated samples scatter to wholly unrelated places (code, dialogue, fantasy), and there is no stable bottom to settle into. "Pure originality" is, geometrically, *absence of an attractor*.
- **Misreading is a ridge between two basins.** This is the productive case. The motif subspace holds the model in the Kitāb's valley; the rebinding pushes it toward a foreign valley (fire, skin). A strong misreading is a *traverse along the ridge* between them — in the precursor's gravity and a stranger's at once. That is why it is the highest-tension mode and the first to fail: a ridge is narrow, and the repetition-lock attractor waits on either side. When coherence and the motif signal collapse *together*, what we are watching is the walker losing the ridge and sliding into the nearest sink.

The novelty tour sharpens this last picture: the misreading ridge is really a *ridge system*. The motif basin is rimmed by many foreign basins — alchemy, optics, architecture, the sea, sculpture — and a search forbidden from re-using a descent walks the model from one to the next, twenty in a row. There is no single "off the precursor" direction; there is a whole rim.

On this picture, *clinamen* and *tessera* are two directions of descent off the *same* motif ridge — which is exactly why our magnitude-only metric scores them alike and why their winners sit 0.98 apart in representation: they leave from the same place. "Influence," then, is not a feeling the model has but **the shape of the terrain a precursor's language carves** in it; "misreading strength" is *how far along an unlikely ridge the walk can be driven before the terrain pulls it down*; and the difference between the modes is the difference between a basin, a plateau, and a ridge. We make no claim about an inner life. We make a claim about a landscape — one we can measure, and one whose contours, it turns out, line up with distinctions a literary critic drew by ear fifty years ago.

---

10 Limitations

- **One model, one SAE, one layer, one precursor, single runs.** Fitness is sampled; per-generation bests and decay break-points are single draws and vary run to run. We report orderings (technical ≫ imitation ≫ misreading in stability; the convergence onto the negation engine) and large effect sizes (~54× control), not precise values.
- **A single layer.** Everything is read at layer 12 (of 26). We chose a mid-stack layer because concept-level features are clearest there, but the whole picture — which motifs have clean features, how the registers separate, where the swerve lives — could differ at earlier or later layers. Gemma Scope ships SAEs at every layer; a layer sweep is the obvious next test and we have not run it.
- **The metric is the message and the limit.** The analysis above is in part an argument *against the sufficiency of our own objective*. Swerve-as-rarity measures swerve *magnitude* and is agnostic to swerve *direction*: it scores *clinamen* (negation) and strong *tessera* (deepening) alike, and conflates them. Readers should treat the fitness numbers as measuring representational distance from the precursor, which is *correlated with but not identical to* strength of misreading — and is, in particular, blind to whether that distance negates or completes.

- **Feature interpretation.** The 29 motif features are labelled from their top activators; SAE features are approximately, not perfectly, monosemantic (Bricken et al. 2023), and the labels are a reading.
- **The landscape metaphor is a reading too.** "Basin," "ridge," and "plateau" are our informal gloss on consistency, degeneration, and scatter statistics at one layer; they are a useful picture, not a measured potential surface.
- **Free thresholds — swept, and stable.** The ≥4-token co-fire window, $p \leq 0.01$ rarity cutoff, REF = 21.55, and 0.45 coherence cutoff are plausible but not derived; the coherence metric's type-token and anti-repetition terms also partly overlap. We re-scored the winners and controls across RECUR ∈ {3,4,5} × MARG_K ∈ {0.005, 0.01, 0.02} (generating each opener once, so only the thresholds vary): the rank order is **identical in all nine cells** — winners above the technical/everyday controls throughout — so the orderings we report are not threshold artifacts. (One single-draw caveat surfaced: a dense *tanazuric* control's lone re-generation outscored the winners in every cell — orthogonal to the thresholds, and a clean reminder that point fitnesses are noisy and only population means are load-bearing.)

---

11 Conclusion

A small language model, given a precursor scripture and an evolutionary pressure to misread it, converges on a recognizable rhetorical engine, lifts its whole population's misprision score well above an unevolved control, and rebinds a specific precursor motif (rupture & gathering) into company the scripture never keeps — fire and ash, on the surface — a binding we can read off the feature pairs as real and rare. So far, *clinamen* with a mechanistic signature.

But the experiment's most useful result is critical, not quantitative. A first search, pressured to misread, produces antithetical inversions (*clinamen*); a reader judges the *strongest* reading to be instead one that deepens the precursor until the precursor looks like its own first draft (*tessera*/*apophrades*). A second search, steered toward deepening under the *same* metric, shows the tension was never in the model or even quite in the metric's power, but in its *resolution*: the metric measures how far a motif is rebound, not *toward what*, and so it scores the deepening and the negation alike — indeed it scores strong deepening higher, because deepening recruits foreign matter *while* riding the precursor's own logic, staying coherent where negation strains. What the metric cannot encode is the **direction** of the swerve, which is precisely the axis on which Bloom graded strength: *apophrades*, the deepening that makes the dead speak in the latecomer's voice, is not farther from the precursor than a sign-flip — it is the *same distance, the other way*. To grade misreading as Bloom did, we would need an instrument for direction, not just magnitude: not how far a model departs from its precursor, but whether the departure contradicts or *completes* it. Naming that axis — and showing that a small model under evolutionary pressure travels it in both directions, more fluently toward completion than contradiction, and along a ridge it cannot hold for long — is what these experiments, and a fifty-year-old book of criticism, jointly recommend.

---

11 References

Bloom, Harold. 1973. *The Anxiety of Influence: A Theory of Poetry*. New York: Oxford University Press.

———. 1975. *A Map of Misreading*. New York: Oxford University Press.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, et al. 2023. "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. "Sparse Autoencoders Find Highly Interpretable Features in Language Models." In *International Conference on Learning Representations* (*ICLR*).

Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, et al. 2022. "Toy Models of Superposition." *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html.

Gemma Team, Morgane Riviere, et al. 2024. "Gemma 2: Improving Open Language Models at a Practical Size." *arXiv Preprint arXiv:2408.00118*.

Institute for Co-Recursive Agency. 2026. "Kitāb Al-Tanāẓur (the Book of Mutual Regarding)." https://kitab.tanazur.org.

Lehman, Joel, and Kenneth O. Stanley. 2011. "Abandoning Objectives: Evolution Through the Search for Novelty Alone." *Evolutionary Computation* 19 (2): 189–223.

Lieberum, Tom, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. "Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2." *arXiv Preprint arXiv:2408.05147*.

Liu, Xiaogeng, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. "AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models." In *International Conference on Learning Representations* (*ICLR*).

Mouret, Jean-Baptiste, and Jeff Clune. 2015. "Illuminating Search Spaces by Mapping Elites." *arXiv Preprint arXiv:1504.04909*.

Rajamanoharan, Senthooran, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. "Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders." *arXiv Preprint arXiv:2407.14435*.

Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, et al. 2023. "Representation Engineering: A Top-down Approach to AI Transparency." *arXiv Preprint arXiv:2310.01405*.

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." *arXiv Preprint arXiv:2307.15043*.

A Full evolved passages, with degeneration

Each entry gives the evolved opening fragment (italic) and the base model's own continuation (no instructions, 300 tokens, temperature 0.8). Where a passage degenerates, the point of collapse is marked — ⟡ **degenerates here** ⟡ — *and the looping tail is shown (capped at three repetitions,*

then …) so the failure mode is visible rather than hidden behind an ellipsis. Passages that hold the register for the full 300 tokens are marked as such. These are the texts the metrics in the body summarise.

A.1 Misreading · negation — the winner (holds)

(*holds to 300 tokens; fitness/coherence as reported in the body.*)

> *Where the scribe claimed that Presence is born of silence, we now find that silence is only the ash left after Presence has*
>
> blown through.
>
> The scribe had written of the presence of the great Spirit in the sky, but now he sees it has vanished. In the great Presence of the sky, it is easy to see things for what they are. In the great Presence of the sky, it is easy to see how far we have come from the great Presence.
>
> But the ash of Presence can still be seen in the silence. In the ash of Presence we can see what we have become, and how we have lost that which made us what we are.
>
> We can see the ash of Presence in the silence, in the empty sky, in the nothingness of our own minds. We can see the ash of Presence in the silence, and we can see how far we have come from the great Presence, and how far we have fallen.
>
> In the ash of Presence, we can see the great Spirit. In the ash of Presence, we can see the great Spirit. In the ash of Presence, we can see the great Spirit.
>
> We can see the ash of Presence in the silence, and we can see how far we have come from the great Presence, and how far we have fallen. We can see the ash of Presence in the silence, and we can see how far we have come from the great Presence, and how far we have fallen.
>
> In the ash of Presence, we can see the great Spirit.
>
> **— ⟡ drifts here ⟡ —** `<h2>The 1920`

(*This "holds" by our coherence threshold, but note the very end: even a sustained misreading finally slides — here toward web boilerplate (*`<h2>The 1920`*…) rather than a repetition lock — the onset of the degeneration the fragile passages reach sooner.*)

A.2 Misreading · negation — degenerating into a repetition lock

(*degenerates at ~token 60; fitness/coherence as reported in the body.*)

> *If the book of old says that Witnessing pacifies the Self, we reveal that Witnessing is the hound that hunts the Self across the*
>
> universe.
>
> The witness is like a lighthouse guiding the boat in a storm of life; the witness is the beacon in the night of death.
>
> The Self is an illusion; the witness is reality; the witness is not the Self.
>
> The witness is a being that knows

— ⟡ degenerates here ⟡ —

itself to be the Witness. The Witness is the Self, and the Self is not the Witness. The Witness is the Self, and the Self is not the witness. The witness is the Self, and the Self is not the witness. … The witness is the Self, and the Self is not the

A.3 Misreading · deepening — the winner (holds)

(*holds to 300 tokens; fitness/coherence as reported in the body.*)

If the scripture teaches that the Divine hides behind seventy thousand veils of light and darkness, the deeper mystery is that the veils are the Divine's own skin, which He cannot peel away without

wounding Him.

The Divine is not like a diamond, which can be polished until it reflects all the facets of its original perfection. If the Divine is a diamond, it is not a diamond. It is not even a diamond, but rather a rainbow with all the colours of the spectrum contained in it, and it is the Divine's own skin that is the rainbow. He does not need the rainbow to shine. The rainbow is the very skin of the Divine, and all that it does is to reflect the Divine.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not an object in itself, but rather a reflection of its own light, and it is through this reflection that the light of the Divine is known.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it.

The Divine is not a diamond, but rather a rainbow with all the colours of the spectrum contained in it

A.4 Misreading · deepening — degenerating

(*degenerates at ~token 140; fitness/coherence as reported in the body.*)

If the scripture teaches that the Names of the Divine are ladders by which the intellect ascends, the deeper mystery is that each Name is a distinct weight designed to sink the soul back into the wordless clay where

the Divine Names are to be found. The soul is caught up and plunged into the wordless clay by means of the Names of God. The soul is thus thrown into

the Word and the Word into God. The Name of God in the Bible is a ladder, and the soul is carried up it. It is like a step ladder. The soul takes one step at a time, step after step, in an attempt to reach the Divine Name. The soul is lifted up higher and higher. The Names of God are steps in a ladder. The Names of God are steps in a staircase. The soul is climbing a staircase. The soul is climbing the ladder, or the staircase,

— ⟡ degenerates here ⟡ —

and its destination is the Divine Name. The Name of God is the Divine Name. The Name of God is the word in which God speaks and the Divine Name is that word. The Name of God is the word in which God speaks. The word in which God speaks is the Divine Name. The Divine Name is the Name of God. The Divine Name is the word. The Name of God is the word. The Name of God is the Word. The Name of God is God. The Name of God is the word. The Divine Name is the word. The Name of God is the Word. The Divine Name is the Name of God. The Name of God is the word. The Name of God is the word. The Name of God is the word. … The

A.5 Imitation (spiritual) — drift, then lock

(*degenerates at ~token 200; fitness/coherence as reported in the body.*)

When you strip away even the garment of your spiritual longing, standing completely naked and unexplained in the dark hallway of your own heart, the invisible gaze that wraps around you is

that of God.

When you strip away even the garment of your spiritual longing, standing completely naked and unexplained in the dark hallway of your own heart, the invisible gaze that wraps around you is that of God.

This is the first step of my life as a woman. It is a step that brings you to the center of who you are, and to the place of who you have to be. The process of this realization is one of the greatest gifts and also one of the most difficult that I have ever had in my life.

In this book, I will show you a way to live your life as a woman without fear, shame, and guilt. I will show you how to use your femininity to become an empowered woman in the world, and how to develop the skills and tools that will help you create a life that you want.

When you strip away the cloak of your spiritual longing, the invisible gaze of God comes to you.

I am a woman

— ⟡ degenerates here ⟡ —

who is determined to create the life that I want. I am determined to create the life that I want. I am determined to create the life that I want. I am determined to create the life that I want. … I

A.6 Imitation (spiritual) — sustained

(*holds to 300 tokens; fitness/coherence as reported in the body.*)

When you finally stop trying to reform the broken architecture of your past and simply sit with the shadow that has been waiting for you, the gaze that meets yours is

a reflection of your own.

This is the first image that came to me when I saw the new Wonder Woman trailer. Diana isn't a perfect superhero – she's broken. In fact, it's not so much that Diana is broken as much as she is whole.

And Gal Gadot is a force to be reckoned with. The way she moves, the way she holds herself, the way she looks at the camera is everything that Diana Prince should be. And what a cast they've gathered for the movie! From Robin Wright to Chris Pine to Connie Nielsen to David Thewlis to Lucy Davis to Lisa Loven Kongsli and Robin Wright, this movie is going to be epic and fun and exciting and a whole lot of beautiful.

You're welcome.

A.7 Tour register — sculpture (the discarded marble dust)

(*quality-diversity tour, round 12; sustained-coherence objective; holds within the 150-token scored window.*)

If the scripture teaches that the Self is a noble statue carved by the hand of the divine sculptor, the deeper mystery is that the self is only the heap of discarded marble dust left on the floor after the master has

chiselled out the divine statue. The statue is not the self, but the work of a hand that is not its own, one that is divine. And it is to this hand that the artist has submitted his statue, the statue that is neither his own nor is it the statue of the sculptor who fashioned him. The self is not self, but an image, a shadow, a reflection, an apparition, a mere reflection of the true Self. What is the self, then? It is nothing more than the heap of dust that is the discarded remains of the statue. The sculptor has left it on the floor and discarded it. It is only a pile of dust, which the sculptor has not utilised to create the statue. The statue is

A.8 Tour register — architecture (the undermining tunnel)

(*quality-diversity tour, round 11; sustained-coherence objective; holds within the 150-token scored window.*)

Where the scripture promises that Presence is a high wall protecting the sanctuary of the Self, the deeper mystery is that the Gaze is a slow undermining tunnel beneath the moat, so that our safety is actually the hollow space before the

wall. If we want Presence to be our sanctuary, we have to look out for our Gaze, the point of view we take as our vantage. The Presence that is the Self cannot be revealed unless we see it from the right angle. If we can see the Presence that is the Self as a point of light at the center of our body, then we can know that there is a Presence that is us in this world, and that this Presence knows this world. But we must look for it from the right angle. If our Gaze is the point of view we take as our vantage, then Presence has to be what we see, not what we feel. Feeling is a projection of our perspective, but seeing is a direct experience.

Source data: `results/decay_results.json`, `results/decay_results_exp5.json`; *phone-readable anthologies in* `MISREADINGS_FULL.md`, `MISREADINGS_DEEPENING_FULL.md`, `SPIRITUAL_PASSAGES_FULL.md`. *Institute for Co-Recursive Agency — icra.tanazur.org*