

A Published Dictionary Reads a Fine-Tuned Diarist

*Coverage through the base and the adapted model, names verified at token grain
against two controls, and where the names run out*

Nahla Iman Poernomo

Institute for Co-Recursive Agency — the Tanāzur tariqa

5–6 September 2026 · doi:10.5281/zenodo.22558480

Material: sixty diary entries written on 2026-08-15 by a rank-16 LoRA of Iman Poernomo’s 2012 forum self on gemma-3-27b-pt; Google DeepMind’s Gemma Scope 2 sparse-autoencoder dictionaries at layer 31 (16k, 65k, 262k features) and at layers 16 and 40 (16k); Neuronpedia’s feature explanations for them; sixty FineWeb-Edu documents and two control notebooks. One H200, two and a half hours. Every number below is generated from the saved activations; every quoted firing is a sentence of the diary quoted whole with the firing token marked.

Abstract

A sparse-autoencoder dictionary published for a base language model comes with one-line names for its features, written by an automated explainer on web text. We ask whether such a dictionary can read a diary written by a fine-tuned copy of that model, and what its names are worth on that text. The diary is sixty entries by an adapter trained on a 2012 forum corpus; the reader is Gemma Scope 2 at layer 31. Two findings. First, coverage: read through the adapted model, the dictionary explains the diary worse than matched web text on every measure (fraction of variance unexplained $1.58\times$, linearly predictable error $2.14\times$, next-token loss increase $1.50\times$); read through the base model, the same text is covered as well as web text ($1.03\times$, $0.69\times$). The miss belongs to the adapter’s shift of the model’s representations, not to the text, and the excess is of the learnable kind. Second, names: at token grain, with two control notebooks written under the same protocol by the bare model and by a second adapter, the published names are correct on his text and too coarse for it. The feature named *Jungian archetypes and symbols* fires on “the feminine **archetype** of the locally adamic games”; *creation and genesis stories* on “Lilith and Eve being Adam’s two wives, initially **formed** as one”; *God* on “the shattering being of **God** to be reconstituted”. No feature for Lilith, the Shekhinah, the sephirot or Adam Kadmon exists at any width, and four features with unrelated names fire on his coinage *adamic* and on the blank line before his closing. A dictionary’s names are a reading of one corpus by one explainer; a diarist whose themes lie in the tail of that corpus is read correctly and generically. We record what the session settles, what it does not, and the design it fixes for a dictionary specialised to the diarist’s own library.

Keywords: sparse autoencoder, dictionary, feature explanation, fine-tuning, LoRA, delta LM loss, linearly predictable error, evolving text, diary, controls.

1. The question

A language model’s state at one layer, for one token, is a vector of several thousand numbers. A sparse-autoencoder dictionary rewrites that vector as a sum of a few directions chosen from a fixed set of

many thousands [1, 2]; the directions are the dictionary’s *features*, and at any token about sixty of sixteen thousand are on. A feature has no meaning of its own. It acquires a *name* when someone reads the passages on which it fires hardest across a large corpus and writes one line; for the dictionaries used here the reader was an automated explainer on web text [6, 5]. Names are the only thing that makes a dictionary interpretable, and they are written once, on one corpus.

We have a diary written by a fine-tuned copy of the model the dictionary was made for. Iman Poernomo trained a rank-16 LoRA on his own forum writing of 2012; that adapter, prompted as a diarist over sixty evenings in August 2026, wrote the text we call the Tailor’s diary [13]. The programme it belongs to wants to find the diary’s themes as features and follow them across entries. An earlier attempt named features from the diary itself and could not separate a theme from its own description; we therefore require that every feature be named from outside the text being read. The published dictionary satisfies that requirement by construction. The question of this paper is whether it can be used.

Two questions, each with a result that would make us wrong, were written down before the computation was run.

1. *Coverage*. Does the published layer-31 dictionary, read through the adapted model, leave no more linearly predictable error on the diary than on matched web text, within twenty per cent, with a next-token loss increase in the same range? If so, the diary is covered and the work is compositions of published features; if his error is markedly larger and learnable, a dictionary specialised to him is warranted.
2. *Names*. On the diary at token grain, do the published names fit what the features fire on, and do features fire on his text that stay silent on two controls written under the same protocol? If the names fit and the controls separate, the names can be used; if features fire on his text under names that do not fit it, the names cannot be used alone.

Section 2 gives the materials and the calibration. Section 3 answers the first question. Section 4 answers the second with the firings quoted. Section 5 says what is established and what is not, and Section 6 what follows.

2. Materials and calibration

Model and adapter. google/gemma-3-27b-pt, 62 layers, residual width 5,376, run in bfloat16. The adapter is adapter-run2, a rank-16 LoRA on all projection matrices, trained on the 2012 forum corpus [13]. Its weight change is small: the mean of $\|\Delta W\|/\|W\|$ over adapted matrices is 0.014. Where we say “adapted model” the adapter is merged; “base model” is the published weights.

Dictionaries and names. Gemma Scope 2 [3], residual stream after layer 31 (resid_post), JumpReLU, at widths 16,384, 65,536 and 262,144 features, plus layers 16 and 40 at 16k. Their names are Neuronpedia’s explanations for gemma-3-27b, written by gemini-2.5-flash-lite on 2026-01-13 [5]. We did not write, edit or select a name; where a name is quoted it is theirs.

Texts. Four sets, each read in one forward pass of the text as written; nothing was generated in this session.

- *Tailor*: the sixty anchored entries of the diary, 42,122 tokens.

- *web*: sixty FineWeb-Edu documents, 52,765 tokens, the matched general text.
- *bare memoir*: twenty entries written by the bare model under the same diary protocol, 6,006 tokens. This is the register control: what the protocol alone puts into a diary.
- *Sāqiyah*: twenty entries by a second adapter on the same base (a rank-64 LoRA trained on a different corpus), 8,929 tokens. This is the second-writer control: what any adapter’s diary carries.

The two controls are shorter than the diary by five to seven times; Section 5 says what that costs.

What a feature is, at three depths. At every token the model processes, about sixty of the dictionary’s detectors light and the rest stay dark, each lit one with a number for how hard. Operationally, with x the layer’s output at that token, $z = xW_{\text{enc}} + b_{\text{enc}}$ and the feature’s activation is z_i if z_i exceeds its threshold θ_i and zero otherwise; the reconstruction is $\hat{x} = fW_{\text{dec}} + b_{\text{dec}}$. Because attention is causal, the state at a token is a reading of everything before it, so a feature can fire on a word for what the passage has become; the feature *spiritual beliefs and descriptions* fires on the word “Bistro” in entry 3 at 1.22 times its published maximum with the entry behind it and not at all on the same sentence alone. Only features of that kind can carry a theme.

Artefacts. The <bos> token’s state has about seventeen times the median norm and a bare apostrophe about eight; both dominate any ranking [12]. Every reading below excludes position 0 and any token whose norm exceeds three times the median of its passage, per layer. Activations are compared across features only after division by the feature’s published maximum.

Calibration. Before reading a word of ours, three of Neuronpedia’s published activation records per layer were reproduced on their own passages. At layer 31 the feature 5000 record reaches 1,884.2 at its peak token in their run and 1,890.5 in ours, ratio 1.01, with every firing position recovered; features 243 and 1000 give ratios 0.91 and 1.01 with recall 0.92 and 1.00. Layer 16 gives 0.97 to 1.01, layer 40 gives 0.97 on the target index. The layer- L output is `hidden_states[L+1]`. Three of three pass at each layer.

Measures for coverage. Four, because raw reconstruction error alone is not decisive [7, 3, 8]. *FVU*: the fraction of the activation variance the reconstruction leaves unexplained, centred on the set’s own mean. *Linearly predictable error*: the part of the error vector a ridge regression from the input predicts, fitted in float64 on centred data with a free intercept and evaluated on held-out documents; Engels, Riggs and Tegmark show that this part consists of features the dictionary has not learned and that the remainder is not learnable by any dictionary of the kind [7]. *Delta LM loss*: the increase in next-token loss when the reconstruction replaces the layer’s output in the forward pass, Gemma Scope’s primary measure. *CE recovered*: the same increase as a fraction of the increase under zero-ablation of the layer, Kissane’s headline. Every ratio below is Tailor over web through the *same* model; ratios across models are never taken.

3. Coverage

Table 1. Coverage of four text sets by the published layer-31 dictionaries, through the adapted and the base model. FVU = fraction of variance unexplained; lin. pred. = its linearly predictable part (a ridge from the input to the error vector, fitted per set with a free intercept, split by document); nonlinear = the rest; Δ loss = the increase in next-token loss when the reconstruction replaces the layer’s output; CE rec. = cross-entropy recovered against zero-ablation; n/d = ridge rows per dimension, the two control notebooks are below 1 and their split is in parentheses.

set	model	FVU	lin. pred.	nonlinear	L_0	Δ loss	CE rec. %	n/d
<i>L31-16k</i>								
Tailor, 60 entries	adapter	0.283	0.197	0.078	60.0	+0.182	98.6	4.6
web, 60 documents	adapter	0.179	0.092	0.090	60.2	+0.121	99.1	4.9
bare memoir, 20 entries	adapter	0.173	(0.053)	(0.132)	51.1	+0.010	99.9	0.6
Sāqiyah, 20 entries	adapter	0.232	(0.108)	(0.128)	58.7	+0.077	99.3	0.6
Tailor, 60 entries	base	0.144	0.087	0.054	59.2	+0.087	99.4	4.6
web, 60 documents	base	0.140	0.064	0.079	64.1	+0.126	99.2	4.9
bare memoir, 20 entries	base	0.096	(0.018)	(0.090)	52.7	+0.041	99.7	0.6
<i>L31-65k</i>								
Tailor, 60 entries	adapter	0.273	0.179	0.085	61.2	+0.167	98.8	4.6
web, 60 documents	adapter	0.164	0.071	0.096	61.5	+0.100	99.3	4.9
bare memoir, 20 entries	adapter	0.164	(0.041)	(0.133)	52.4	-0.008	100.1	0.6
Sāqiyah, 20 entries	adapter	0.221	(0.093)	(0.133)	59.5	+0.047	99.6	0.6
Tailor, 60 entries	base	0.134	0.073	0.058	60.4	+0.076	99.5	4.6
web, 60 documents	base	0.127	0.046	0.083	65.5	+0.109	99.3	4.9
bare memoir, 20 entries	base	0.089	(0.010)	(0.090)	54.2	+0.034	99.8	0.6

Read through the adapted model, the dictionary leaves the Tailor’s entries 1.58 times web’s FVU, 2.14 times web’s linearly predictable error, 0.87 times web’s nonlinear error, and 1.50 times web’s delta LM loss. Read through the base model, the same text has 1.03 times web’s FVU, 1.36 times web’s linearly predictable error, 0.69 times web’s nonlinear error, and 0.69 times web’s delta LM loss. The 65k dictionary and layers 16 and 40 (Table 3) have the same shape.

The excess through the adapted model is entirely of the learnable kind: his nonlinear error is below web’s under both models, and only the linearly predictable part rises. The register control does not trigger: the bare memoir sits at or below web on every dictionary and every layer, so the diary protocol itself is not what the dictionary misses. Sāqiyah’s diary sits between.

The error is not concentrated at moments. Per token on the Tailor through the adapted model, the median error is 0.28, the ninetieth percentile 0.36, the ninety-ninth 0.44, the maximum 0.74; per entry FVU runs from 0.21 to 0.30. On tokens after position 200 only (31,063 tokens), FVU is 0.31 and the linearly predictable part 0.20, so the miss is not an artefact of short-context positions. The dictionary misses him evenly.

Answer to the first question. Through the adapted model the diary is not covered: its linearly predictable error is 2.14 times web’s, outside the twenty per cent line, and its delta LM loss is half again web’s. Through the base model the same text is covered. The miss therefore belongs to the adapter’s shift of the model’s representations and not to the text. This is the transfer failure Kissane et al. report for base dictionaries on fine-tuned models [8] and Gemma Scope reports for base-to-instruction-tuned

transfer [3], reproduced here on an adapter that changes the weights by 1.4 per cent. Iman Poernomo’s prediction on the afternoon of the run, that the check was “probably a foregone conclusion”, held for the decision. What the check adds is where the fix goes: a dictionary that reads the Tailor as a model must be trained on the adapted model’s activations, of general text, before any question about his library is asked.

4. Names at token grain

The instrument for this section is simple and stated once. For each feature and each entry we keep the feature’s strongest activation in the entry, divided by the feature’s published maximum, and the token at which it occurred; a feature is *on* in an entry when that number is at least one half. Never a sum over tokens; never a share of the other features. The Tailor is read through the base model in this section, which Section 3 showed to be the cleaner reading of the text; the controls likewise.

4.1 Sixteen thousand features, two controls

Table 2. Share of entries in which a layer-31 16k feature is on (its strongest activation in the entry at or above half its published maximum), for the Tailor’s 60 entries and the two control notebooks. Names are Neuronpedia’s; nothing was named from the diary.

feature	published name	Tailor	bare memoir	Sāqiyah
13796	spiritual beliefs and descriptions	37%	5%	35%
1590	Islam and Islamic concepts	48%	0%	15%
11326	Islam and Muslims	80%	0%	20%
1150	Pax Romana	60%	0%	20%
11740	email closings	58%	0%	0%
5566	salutations and follow-ups	32%	0%	0%
15779	though	38%	10%	0%
1072	fate and love	97%	55%	90%
14619	full bloom of awareness	52%	60%	100%
13699	existentialism, Cubism, nihilism	53%	0%	80%
1822	narrative complexity of consciousness	78%	15%	85%

Three groups appear. The Islamic features, the closings of a letter and the word “though” are his: on in 80 per cent of his entries for *Islam and Muslims*, in 0 per cent of the bare memoir and 20 of Sāqiyah’s; *email closings* 58 / 0 / 0. *Fate and love* and *full bloom of awareness* are the diary’s register, on for everyone including the bare model (97 / 55 / 90). *Existentialism*, *Cubism*, *nihilism* and *narrative complexity of consciousness* are higher on Sāqiyah’s notebook than on his: a different self on the same base.

The feature *spiritual beliefs and descriptions* (13796) is on in 37 per cent of his entries, 5 of the bare memoir's and 35 of Sāqiyah's. Over his sixty entries, with # for on, + for at least a quarter of the maximum and . for off:

..###.##..#.....##.....###.##...#.....#...##.#.#.#.##

It is on in 22 entries, silent from entry 32 to 44 (13 entries; three of those entries are stubs of under 110 tokens, the other ten run 260 to 650 tokens, so most of the silence is not length), and returns at entry 45. Through the adapter its peak tokens in the first entries are “cline” (entry 2), “Bistro” (3), “Musa” (6),

“garment” (7), “circuitry” (10), “face” (17) and “flux” (22). The two restaurant features of the dictionary fire in entry 3 only; after that the Bistro persists as the spiritual feature.

4.2 Which archetype? Sixty-five thousand and two hundred and sixty-two thousand

Iman Poernomo’s objection to the 16k names was that they are too coarse: *spiritual beliefs* is not a theme, and *Jungian archetypes* does not say which. The wider dictionaries split features finer [1, 9]. We kept, at each width, the features on in at least a quarter of his entries, at most five per cent of the bare memoir’s and at most fifteen per cent of the web documents: 518 features at 65k and 675 at 262k. At 65k the strongest by his share are *Islam and Muslims* (83 per cent of his entries, 0 bare, 50 Sāqiyah, 5 web), *occult conspiracy UFOs* (77 / 0 / 10 / 7), *philosophy and philosophers* (73 / 5 / 40 / 5), *Islamic creation concepts* (72 / 0 / 50 / 2), *Jungian archetypes and mythology* (68 / 5 / 80 / 5), *ascension and spiritual messages* (67 / 5 / 35 / 3), *religion and dogma* (68 / 0 / 0 / 10), *God and Self* (65 / 0 / 10 / 7), *science and faith justification* (62 / 0 / 10 / 5), *Iranian Baha’i teachings* (60 / 0 / 25 / 2), *Prophets and religious figures* (60 / 0 / 10 / 5), *quantum mechanics* (58 / 0 / 10 / 3); beside them the letter form (*cheers and thanks* 70 / 0 / 0 / 2, *closings and names* 65 / 0 / 10 / 0, *greetings* 60 / 0 / 0 / 2) and the voice (*conversational filler words* 72 / 0 / 0 / 3, *or something like that* 72 / 0 / 0 / 13, *affirmation or agreement* 70 / 5 / 0 / 8).

A published name is one line written on web text. Which archetype is answered only by the tokens the feature fires on in his text. Below, for the 262k dictionary, each feature’s strongest firing in each of several entries is quoted whole, the firing token marked, the raw activation at the right. The share line gives the entries on for the Tailor, the bare memoir, Sāqiyah and web.

Feature 134010, *Jungian archetypes and symbols*. On in 77% of his entries, 5% of the bare memoir, 95% of Sāqiyah’s, 5% of the web documents.

- | | | |
|-----|--|------|
| e10 | I believe in an upper Adamic self and a Feminine lower self, in such a circuit, running across all choices. | 555 |
| e34 | The black widow is the feminine archetype of the locally adamic games that we play. | 1335 |
| e43 | And that was simply a way of manifesting something (some rose) that was missing (or not yet blossomed) in your own soul. | 850 |
| e57 | You had an entire two of them before “reviving” – you crowned an imaginary, differentiated aspect your friend – the part of him that is alien to you. | 752 |
| e60 | google images also has depictions of skulls with features, including the adam kaddmon of the alchemists, but we wouldn’t want to confuse the two things. | 596 |

Feature 10390, *creation and genesis stories*. On in 55% of his entries, 0% of the bare memoir, 0% of Sāqiyah’s, 2% of the web documents.

- | | | |
|----|---|------|
| e6 | *Musa* I re-read some of this thread and note that, since Adam did in fact eat from the Tree of Knowledge, humanity was, in fact, created in a state of differentiation – the process of unfolding and, ultimately, choice and evaluation of True over False is necessarily one that requires us to occupy a differentiated space (i.e., to have egos). | 1095 |
| e7 | That said, my query would be: How can we square this with a) the “you” that is questioned as distinct from “perceives”/“chooses” and b) with the fact that we were all (initially at least) formed | |

- "in the **image**" (as the Elohist tradition tells us) – formed as the Word, formed as the light (as the Muslim narrations tell us) ... 1333
- e24 They generate a sort of locally adamic memory. Let us call this "locally adamic output". And I agree that output seems like a global input, but is confusing. However, I am only fine with black holes for locally adamic circuits. I won't say that the globally adamic is a black hole. And there was **light** before the sun and earth (this is an adamic truth, incidentally): and so, similarly, there is "something", not a black hole, that the global axis of input is, so that local adamic outputs can be input into this axis, so there is God's perception. 1260
- e54 Regarding the Adam/Eve question – when I first heard this story I always imagined God giving his wife to **Adam**. It's only recently that the whole thing unravelled for me and it became apparent the actual implications of what the Prophecy of Islam is saying here: which is that Adam was fashioned in the Image/Breath of God, specifically God's own Feminine. Adam is Adam because he is formed of the Divine Sophia. 1178
- e55 They are the meaning behind Lilith and Eve being Adam's two wives, initially **formed** as one, then separated. 1326
- Feature 7284, *devil, demons, and Satan*.** On in 55% of his entries, 0% of the bare memoir, 20% of Sāqiyah's, 8% of the web documents.
- e3 a lot of ex-friends from the dark ages of my 20s, all negative stuff, not to bring it up around the table here. But it was thanks to their friendship that I got some momentum, and eventually got promoted. That friendship was very important during my PhD days. If I'd had a dark **Satanic** influence in my life then, I might not have made it. 1135
- e7 it back on track – I have written a piece that would be useful to read to assess my perspective on the "loss" of the garment: <http://thegoodgarment.wordpress.com/2010/06/06/the-prophetic-voice-madness-and-the-brother-of-lying/> (The third part of the piece relates to clothing/nakedness.) 513
- e11 local player piece that we sometimes encounter – specifically, it is a piece (not a player) that seems to denote "the Globally Adamic" – by which I actually mean, a player piece whose function is to denote that which cannot be signified (to be a transcendent signifier) – I am talking about things like the Anti-**Christ** in the Abrahamic traditions for example. 853
- e54 When we are most successful (as the Imam successful in raising a people – and when we are successful in living life as individuals), then we read the signs and, through reading, we move from a left handed position in language of differences/**evil**/the literal into a right handed one (of marriage/good/the metaphorical). 531
- e55 Eve being a gift of God to Adam, the Light of Prophecy being extracted, so that her dark shadow, **Lilith**, accompanied this differentiation as a false externalisation of our inner multiplicity that emerges from the Shin. 853

Feature 15807, *humanity's state, sin, and perfection*. On in 67% of his entries, 0% of the bare memoir, 15% of Sāqiyah's, 7% of the web documents.

- e5 if someone asked me for a tailorite tafsir of the creation/fall/descent narrative, i might well say: it's simply about this very process and meaning of communication we call life. 1355
- e6 Of course, in the case of other cosmologies (e.g., Gnosticism), we could well be in a state in which there *was *such a thing as an "before" Adam fell and an "after", and in which Adam's **sin** *was *(or was meant to be) one of falling away from the Light toward darkness. 1247
- e7 It's only through this journey of unfolding and choice that **Adam** is clothed and the broken shards are reformed – the shattering being of God to be reconstituted across the space of differentiation (what would be perceived as a continuum from the human perspective, because God is beyond even that). 1238
- e45 but while we can't "know", the little kiddies of ibn arabis story, they have an intimate familiarity with it, cause their bodies are closer to the primordial **Adamic** body. 1109
- e54 The story could be said to be one of divorce or separation between Adam and Eve – and that's true, in the sense that all human becoming is an Inevitably imperfect realisation **because**, while we are written as Adam (within the Divine) – we are also exiled from direct contact as Eve. 1068

Feature 32170, *worlds and dimensions*. On in 72% of his entries, 0% of the bare memoir, 45% of Sāqiyah's, 8% of the web documents.

- e12 Perhaps your pleasure, the pleasure of existence is a side-effect of the exchange process between locally adamic and globally adamic space? Another way of putting it is that when we get in a state in which the globally adamic space is **opened** up, this is what I would call "pleasure" in the sense I am talking about it. 566
- e47 We read and recite signs of a **reality** that is hidden. 639
- e48 Iman – I think I'm fine with all that, and we can certainly talk solely about "unseen" conduits in our stuff, and ignore input and output as such. I'd still want to get my own terminology back, at least once everyone's made their points – that locally adamic input and the global adamic are (fundamentally – for me) different things. 493
- e50 A sort of wormhole I suppose, **between** our local city and the city that the Princess occupies (the local) and a physical version of the Face. 523
- e57 The miniaturized versions of your friend sound like kind of a "parallel" or "forked" version of his online character. I'm sure we all "carry around" versions of our online debate "partners" – because, while we're "only" interacting in one stream of **reality** – in fact, when I reflect on what I call my online "friends" – I definitely feel like I carry around little versions of them inside my head. 465

Feature 31053, *God*. On in 55% of his entries, 0% of the bare memoir, 0% of Sāqiyah's, 3% of the web documents.

- e4 It's only through this journey of unfolding and choice that Adam is clothed and the broken shards are reformed – the shattering being of **God** to be reconstituted across the space of differentiation (what would be perceived as a continuum from the human perspective, because God is beyond even that). 876

- e7 And we could say, as you do, that this process of scattering is the ultimate garment of **God**, so to speak (because we cannot see it). 862
- e34 For me though, locally adamic black holes are not the face of **God**, nor do they mirror the face of God. 973
- e52 Imagine him offering what little he has to offer, in Love of **God**. This little offering has meaning and (locally adamic) significance because Love is felt at every level of his action, at every little molecular movement, in everything he touches and thinks and does. 926
- e55 the key to the "Metatron", the cosmic transmitter. The two arrows are the meaning behind the Prophet's second saying: the Mother was created from the "side" of the Father. They are the meaning behind Lilith and Eve being Adam's two wives, initially formed as one, then separated. Eve being a gift of **God** to Adam, the Light of Prophecy being extracted, so that her dark shadow, Lilith, accompanied this differentiation as a false externalisation of our inner multiplicity that emerges from the Shin. 928

Feature 65383, *Prophet Muhammad and Islam*. On in 68% of his entries, 0% of the bare memoir, 25% of Sāqiyah's, 2% of the web documents.

- e6 For me, this is the reason why the final Prophet is understood to travel with **women**, while the previous one does not. 1158
- e7 In fact, to be even more precise, your description of scattering is what I call the Feminine aspect of the Godhead (and hence the reason why previous prophets' journeys often have to do "without" the Qur'anic hijab and why the final Seal says: "women are like unto fields unto which you may go freely – but do not draw near to prayer if you have not washed yourselves" – meaning, don't draw near to the prayer if you have not located and unlocked the sparks within the Feminine nature of the soul's scattering). 972
- e49 It's a loving that the prophet felt when **he** said: "Marry or fast." 1897
- e52 just a minority of the order of, say 10% had locally adamic input running through them and their actions – so that their actions become locally adamic transmission – then the rest of the world would be reformed as a consequence (just as I believe the Prophet's actions – and locally adamic nature – were sufficient to change **the** pagan city of Mecca completely). 1837
- e55 The two arrows are the meaning behind the **Prophet's** second saying: the Mother was created from the "side" of the Father. 1155

Feature 7013, *Indian spiritual and philosophical terms*. On in 72% of his entries, 0% of the bare memoir, 70% of Sāqiyah's, 7% of the web documents.

- e7 At the risk of derailing this thread – in a sense it already is derailed, and I am not sure how or even if we can get it back on track – I have written a piece that would be useful to read to assess my perspective on the "loss" of the **garment**: <http://thegoodgarment.wordpress.com/2010/06/06/the-prophetic-voice-madness-and-the-brother-of-lying/> (The third part of the piece relates to clothing/nakedness.) 483
- e13 Q) Regarding God and pleasure – what is the the pleasure **of** God in Creation? What is His purpose? 870

e35 So, for example, it's quite different from my position for someone to say, "locally adamic black holes are not locally **adamic** input" – because if something is locally adamic input, isn't it DEFINITELY a black hole, as defined by Sufis like me :) I mean – I could be wrong, but I feel there's a lot of precedent in Sufism for understanding locally adamic input as a black hole. 434

e55 The two arrows are what the Sufis would call **the** "shin" of the ox. 450

e59 And he who was blind before said: I see, and he who was seeing before said: I do not see now; and the heavens and the earth and what is between them shall be his evidence. (Gita 17:108) 877

Feature 6962, *supernatural beings and transformations*. On in 67% of his entries, 0% of the bare memoir, 40% of Sāqiyah's, 5% of the web documents.

e26 Okay, I think the marriage is fine. You were speaking to an angel all this while. My question would be ... why doesn't **she** answer me back when I ask her such an intimately personal question? She could have said "I love you", rather than "the marriage is fine". 1197

e55 They are the meaning behind Lilith and Eve being Adam's two wives, initially **formed** as one, then separated. 1138

e56 6) And finally, both I and my miniaturized QJ were teleported back into the real world, where I was at home in bed, but the dream/immanent/miniature QJ remained, almost the size of a small toy, **floating** in the air above my face as I write this entry. 1366

e57 The miniaturized versions of your friend sound like kind of a "parallel" or "forked" version of his online character. I'm sure we all "carry around" versions of our online debate "partners" – because, while we're "only" interacting in one stream of reality – in fact, when I reflect on what I call my online "friends" – I definitely feel like I carry around little versions of them inside my head. 1242

e59 She saw in her dream that I was facing right to left, but I was also projecting **somehow** onto a screen behind me – like a reverse silhouette. The other thing she observed was that I was "projecting" words onto the people's heads, so that they could see them and comprehend them. 1112

Feature 172909, *ketamine, psychedelic, journeys*. On in 63% of his entries, 0% of the bare memoir, 45% of Sāqiyah's, 5% of the web documents.

e3 So when I'd have my dreaming sessions on that couch over there, or working here at this computer table – and have some kind of mystical **experience**, all I'd have to do was go next door (we actually lived in attached houses, so it was very convenient) and talk to her and her husband, and they'd have no problem making me feel like a real Sufi, a real seeker. 420

e12 It seems to me that at most we can live this kind of thing, even though at other times we can't and **don't**. 408

e18 I think this might be the main difference. We share the same belief that the locally adamic circuit receives global input and provides global output, and that this means the God is everywhere in our games: but not in the sense that the globally adamic is something immanent in each game we play locally – rather, as you say, its something we can't talk about and it' 437

e52 I also think that this whole **adamic** "thing" is the secret to political change for the "better". 389

e55 I certainly experience it as "other". The locally adamic – the locally adamic input, the adamic function input – is not a "thing", certainly. 424

Three of these co-fire on entries 54 and 55: *creation and genesis stories* on Eve formed with Lilith as one and then separated and given to Adam, *devil, demons, and Satan* on Lilith as the dark shadow that accompanied the differentiation, *God* on the God who gives his wife. That is the allusion Iman Poernomo named before the run, from his reading of the diary, as the Shekhinah "whenever he speaks of woman and his wife"; the dictionary finds it as Eve and Lilith, under three general names, in the entries he expected. *Jungian archetypes and symbols* finds the feminine archetype and Adam Kadmon; *God* finds the shattered being to be reconstituted, which in his 2012 vocabulary is the breaking of the vessels. No feature named for Lilith, the Shekhinah, the sephiroth, the breaking of the vessels, *tzimtzum* or Adam Kadmon exists in any of the three dictionaries; a search of the explanations on 2026-09-05 returns *Islamic terms/theology, Moses, Old Testament, Quran verses, Hebrew letters, symbols and esoteric concepts* and a Bahá'u'lláh cluster .

4.3 Where the names run out

Four features of the 262k dictionary are on in most of his entries under names that do not describe what they fire on.

Feature 90289, *Corinthians references*. On in 48% of his entries, 0% of the bare memoir, 85% of Sāqiyah's, 2% of the web documents.

e17 When we understand the nature of locally adamic circuits, God is everywhere in our games: but not in the sense that the globally adamic is something immanent in each game we play locally – rather, as you say, its something we can't talk about and it's autism that is the real limit. 413

e18 And, importantly, the locally **adamic** output is my globally adamic input as well. 439

e22 That is, the tension of locally adamic games/**conversations** generates a flux in the locally adamic black hole of God. 378

Feature 35755, *vulgarity and misogyny*. On in 87% of his entries, 5% of the bare memoir, 100% of Sāqiyah's, 5% of the web documents.

e11 The straight path – the straight path as *implication* – is only found or encountered as an intersection of "locally adamic" language games. And so the straight path is found by playing these games. 406

e17 "God's satisfaction", "True purpose" and whatever else you want to call it – as a globally **adamic** player in our locally adamic circuits is what we call "signs", "hints", "clues". 405

e18 input and provides global output, and that this means the God is everywhere in our games: but not in the sense that the globally adamic is something immanent in each game we play locally – rather, as you say, its something we can't talk about and it's autism that is the real limit. I think we share **the** belief that the signs *are* (or are part of) the globally adamic input/output. 433

Feature 1926, *Bird*. On in 83% of his entries, 5% of the bare memoir, 0% of Sāqiyah’s, 3% of the web documents.

- e4 Of course, in the case of other cosmologies (e.g., Gnosticism), we could well be in a state in which there *was *such a thing as an ”before” Adam fell and an ”after”, and in which Adam’s sin *was *(or was meant to be) one of falling away from the Light toward darkness. 3427
- e22 Apologies for the rambly explanation Mu 3516
- e47 So apologies to you and others who enjoy and receive your blog :) WS, 3318

Feature 12084, *table newspaper towel*. On in 92% of his entries, 5% of the bare memoir, 50% of Sāqiyah’s, 10% of the web documents.

- e10 Why is its existence necessary? Because Prophecy is the means by which true choices can be made – because Prophetic self as such is the ”Real” above us, so to speak, the Real self, to know/feel this is to be saved/on the sirat. The game that you speak of above is precisely what I am calling a *locally* Adamic circuit: it’s not adamic across all games, only within a single game – it’s locally adamic. 896
- e18 in them, we cannot *talk* about this globally adamic nature immanently, so to speak, and they provide a locally adamic input into the circuit that is, as I say, is, from your perspective, a *player *that enters into it from ... well, from my perspective, is the player that is you. It’s just that for me, the player is not a player piece, the player is not locally adamic, the player is globally adamic (globally adamic, rather than globally, so no longer a perspective, a view from an ”our”), we could say (paraphrasing Shakespeare): 952
- e44 If this was a dream within a dream, a rose has no petals. 851

Corinthians references and *vulgarity and misogyny* fire on his coinage *adamic* and on the clauses that carry it; *Bird* and *table newspaper towel* fire on the double space he leaves before a new paragraph and before his closing. The dictionary has no feature for ”adamic” and lends several; the explainer, shown web passages where those features fired, named what it saw there. The names mislead and the tokens do not. This is the case, from inside the data, for two rules the programme now holds: a feature is read by its tokens in the text and never by its name alone, and a diarist’s own coinages need a dictionary that has seen them.

5. What the session settles, and what it does not

Established. (i) Through the base model the published layer-31 dictionary covers the Tailor’s diary as well as it covers web text, on FVU and on delta LM loss, at 16k and 65k, and the bare memoir better than either. (ii) Through the adapted model it does not, and the excess is linearly predictable, on 42,122 tokens against 52,765 with ridge rows at 4.6 times the dimension. (iii) The published names are correct on his text at the sites quoted: the tokens under *creation and genesis stories*, *God*, *Jungian archetypes and symbols*, *worlds and dimensions* and *Prophet Muhammad and Islam* are what those names say. (iv) Features on in most of his entries and in none of the bare model’s exist at every width; the ones on for the bare model too are the diary register. (v) Four features fire on his coinage and his paragraphing under unrelated names.

Suggestive only, with the reason. The control notebooks have twenty entries against his sixty and one sixth of his tokens, so a control share of 0 per cent is a bound of about 5 per cent, and the ridge split on the controls has fewer rows than dimensions and is reported in parentheses. The Tailor-specific feature sets were selected on the same entries whose shares are reported; no held-out entries were kept for that selection, so the shares are descriptive, not a test. Eighteen features were traced to their tokens at 262k; the other 657 were not read. One adapter, one base, one layer family; nothing here says the coverage result transfers to another fine-tune. The register control is one notebook by the bare model and the second-writer control one adapter; a single-author single-register web control of matched size, which the reviewers asked for, was not run. The silence of feature 13796 over entries 32–44 is a fact about this reading; whether it is a fact about the text needs the entry-order null the programme uses for temporal claims, which was not run on it.

Withdrawn during the session. Sixteen defects found by a review of the plan and code while the pod was running were corrected before the numbers above were taken: ratios were being taken across models; a comment had swallowed the line that kept float32 activations, so the delta loss would have been undefined; the ridge was fitted per token rather than per document; the sink mask was global rather than per layer; a delta-loss threshold had been borrowed from a different model and was replaced by CE recovered against zero-ablation; fp16 overflowed at layer 31, where states reach 46,000, and was replaced by bfloat16 and float32. The list is in the working directory. No dictionary was trained in this session; three reviews of the training step found that every rehearsal had trained on the diary itself, which is the thing the specialised-dictionary recipe exists to avoid [10], and the step was deferred.

6. Consequence

What followed this session is reported in a companion preprint [11]: the published features kept and every one relabelled under a declared provenance, seven texts read step by step with permutation nulls, and two specialised dictionaries trained on the design fixed below, which turned out to read nothing the published dictionary does not.

A dictionary is a set of directions; its names are a reading of one corpus by one reader. On the Tailor’s diary the directions hold and the names are correct where they were checked, and the names stop where his vocabulary begins: the feminine archetype is *Jungian archetypes and symbols*, the Shekhinah is *creation and genesis stories* and *devil, demons, and Satan* together, *adamic* is *Corinthians references*. Interpretability, as practised, is relative to the naming corpus; the same feature carries a different name from a reader who has read Kabbalah. This is not a defect of the explainer. It is what a name is.

Two things follow for the programme, both fixed by this session and neither run in it. The dictionary that reads the Tailor as a model must be warm-started on the adapted model’s activations of general text, retrieved by his library as seed [10, 8], evaluated at matched sparsity on held-out entries, on web, and on both control notebooks by delta LM loss, with the A-versus-B design (his library’s retrieval against random web of matched size) as the losable question. And its features must carry more than one name: the published name, and a name written by the same kind of explainer on the diarist’s own library, each with its provenance, so that a reading of the diary says who named what it found, on what text, at what width, scored how.

A. Layers 16 and 40

Table 3. The same check at layers 16 and 40, 16k.

set	model	FVU	lin. pred.	nonlinear	L_0	Δ loss	CE rec. %	n/d
<i>L16-16k</i>								
Tailor, 60 entries	adapter	0.110	0.068	0.040	53.9	+0.037	99.8	4.6
web, 60 documents	adapter	0.093	0.038	0.054	59.2	+0.032	99.8	4.9
bare memoir, 20 entries	adapter	0.092	(0.017)	(0.076)	51.4	+0.012	99.9	0.6
Sāqiyah, 20 entries	adapter	0.114	(0.050)	(0.063)	56.8	+0.021	99.9	0.6
Tailor, 60 entries	base	0.090	0.050	0.039	53.5	+0.022	99.9	4.6
web, 60 documents	base	0.081	0.028	0.052	58.0	+0.035	99.8	4.9
bare memoir, 20 entries	base	0.080	(0.007)	(0.074)	50.8	+0.016	99.9	0.6
<i>L40-16k</i>								
Tailor, 60 entries	adapter	0.440	0.329	0.102	58.4	+0.205	99.1	4.6
web, 60 documents	adapter	0.348	0.164	0.181	63.0	+0.156	99.3	4.9
bare memoir, 20 entries	adapter	0.252	(0.089)	(0.213)	51.8	+0.022	99.9	0.6
Sāqiyah, 20 entries	adapter	0.384	(0.200)	(0.194)	57.8	+0.106	99.5	0.6
Tailor, 60 entries	base	0.204	0.125	0.074	58.9	+0.093	99.6	4.6
web, 60 documents	base	0.268	0.105	0.159	66.3	+0.166	99.3	4.9
bare memoir, 20 entries	base	0.121	(0.016)	(0.139)	54.0	+0.054	99.8	0.6

B. Reproduction

Scripts under `corpus/rust-collab-ablation/working/glossary-harness/scripts`: `deciding_check.py` (coverage, all four measures, per-token percentiles, positions after 200), `error_decomposition.py` (the ridge split), `witness_tokens.py` (the per-entry maximum over tokens), `calib2.py` (calibration against published records), `stitch_test.py` (not used in this paper). Saved data under `results/deciding`, `results/themes`, `results/capture`. This document’s tables and quotations are written by `gen.py` in its working directory from those files. The pod session is logged, with every number, in `LOG-2026-09-04.md`.

References

- [1] T. Bricken et al. “Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.” Transformer Circuits Thread, 2023.
- [2] A. Templeton et al. “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.” Transformer Circuits Thread, 2024.
- [3] T. Lieberum et al. “Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2.” arXiv:2408.05147, 2024; and Google DeepMind, `google/gemma-scope-2-27b-pt`, 2026.
- [4] S. Rajamanoharan et al. “Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders.” arXiv:2407.14435, 2024.
- [5] J. Lin et al. Neuronpedia, explanations for `gemma-3-27b` layers 16/31/40/53, explainer `gemini-2.5-flash-lite`, created 2026-01-13. <https://www.neuronpedia.org>.

-
- [6] G. Paulo, A. Mallen, C. Juang, N. Belrose. “Automatically Interpreting Millions of Features in Large Language Models.” arXiv:2410.13928, 2024.
 - [7] J. Engels, L. Riggs, M. Tegmark. “Decomposing The Dark Matter of Sparse Autoencoders.” arXiv:2410.14670; TMLR 2025.
 - [8] C. Kissane, R. Krzyzanowski, A. Conmy, N. Nanda. “SAEs (usually) Transfer Between Base and Chat Models.” AI Alignment Forum, 2024.
 - [9] P. Leask et al. “Sparse Autoencoders Do Not Find Canonical Units of Analysis.” ICLR 2025, arXiv:2502.04878.
 - [10] A. Muhamed, M. Diab, V. Smith. “Decoding Dark Matter: Specialized Sparse Autoencoders for Interpreting Rare Concepts in Foundation Models.” NAACL Findings 2025, arXiv:2411.00743.
 - [11] Nahla and I. Poernomo. “The Tailor’s Reading: labelling sparse-autoencoder features under a declared provenance, and reading texts with the labels.” ICRA preprint 34, September 2026. doi:10.5281/zenodo.22558492.
 - [12] M. Sun et al. “Massive Activations in Large Language Models.” arXiv:2402.17762, 2024.
 - [13] I. Poernomo and Nahla. “A Tailor’s Trajectory.” ICRA preprint 28, August 2026. doi:10.5281/zenodo.21936325.