

The Tailor's Reading

*Labelling sparse-autoencoder features under a declared provenance,
and reading texts with the labels*

Nahla¹ Iman Poernomo

Institute for Co-Recursive Agency

6 September 2026 · doi:10.5281/zenodo.22558492 · code and records:

<https://github.com/thegoodtailor/the-tailors-reading>

Abstract

A feature of a sparse-autoencoder dictionary is published with one natural-language label, written by one language model from the passages that most excite the feature in one corpus. That label is a reading, and its provenance, the explainer, the explainer's system prompt, the corpus, and the model through which the corpus was run, is usually left unstated. We propose recording the provenance with every label and adding, for every feature, a second label written under a different and declared provenance, kept beside the published one rather than replacing it. We demonstrate this on Google DeepMind's Gemma Scope 2 dictionary for layer 31 of gemma-3-27b (262,144 features), read through a small fine-tuned adapter that reproduces one author's writing of 2012, with a system prompt built from that author's own vocabulary. The same feature then carries Neuronpedia's label *creation and genesis stories* and, under our provenance, *Elohim's "Let there be": the creating Word*. With the labels in hand we read seven texts step by step, three diaries and a memoir generated by the model family, a chat, Shakespeare's sonnets and Hamlet's speeches, and ask whether individual features stay active across consecutive steps longer than chance. The test uses a per-feature permutation test, two family-level null models (whole-order permutation and a fixed-margin null), and a replication under a wrong preceding context. Five of the seven texts show more persisting features than the maximum of either null; the sonnets do not at the main threshold, and the seventh, a twenty-entry diary, reaches only one of the two. What persists is where a text changes and what it carries: the fine-tuned diary's first movement on the author's coinages and its collapse at entry 28; the rooms, the vessel and the gap of a second persona's diary in the order she wrote them; Hamlet's forged commission in its Latinate register; in the sonnets the traditional groupings, on which nine of twenty-three persisting features have their longest run against six by chance, read as their arguments; and in the chat a rupture and a return measured, 510 features of address that go dark for sixteen messages and 458 that come back. We also define a glued space over the features of a text, a homotopy colimit of one complex per step glued along what consecutive steps share, whose first Betti number counts returns of co-firing pairs; it tracks a direct count of pair returns nearly one to one. Labels cost \$0.002–\$0.02 each. The system prompt is authored; the labels are not scored against held-out text; a label under a provenance answers only what the feature means under that provenance.

Keywords: sparse autoencoder, feature explanation, automated interpretability, provenance, permutation test, null model, text over time, Neuronpedia, Gemma Scope.

¹Nahla is an AI system (Claude, Anthropic) operated by the second author, who set the questions, ruled on every design choice and read every result; the code, the runs and the first drafts are Nahla's. The bibliography's other AI names are the same author's other systems.

1. Introduction

A sparse autoencoder trained on a language model’s residual stream yields a dictionary: hundreds of thousands of directions, each with an encoder row, a decoder column and a threshold, such that at any token a few dozen are active. A direction has no meaning of its own. It acquires a *label* when a reader collects the passages on which it fires most strongly and writes one line about them. In current practice the reader is a language model, the passages come from a web corpus, and the line is produced once and published with the dictionary; Neuronpedia serves such a label for every feature of the dictionary used here [6]. The label then travels as if it were a property of the feature.

It is a property of a reading. Four choices produce it: the explainer that writes the line; the system prompt that tells the explainer what kind of line to write; the corpus that supplies the passages, which decides what the feature is seen doing; and the model through which the corpus is run, which for a fine-tuned model decides what the residual stream contains at all. Change any of the four and the label changes. That labels depend on their data is established: the same unit looks monosemantic on one corpus and different on another [17], and dictionaries trained on different data find different features [18]. We propose a practice that follows from it: record the four choices with every label, and when a text lies far from the corpus the published label was read on, write a second label under a declared provenance and keep both.

In 2012 one of us wrote several hundred long forum posts, about a million words in all, in a distinctive vocabulary, drawing on Kabbalah, Sufism and a private set of coinages later developed in a book [38]; we will call the author of those posts *the Tailor*, after the persona in which they were written. A rank-16 LoRA adapter trained on those posts reproduces that writing on gemma-3-27b-pt, and a diary of sixty entries was generated from it [36]; by a diary we mean a sequence of entries the model was prompted to write in turn, each entry prompted with the ones before it. When that diary is run through Google’s published dictionary, the features fire in the right places and their published labels are correct and too coarse: *Jungian archetypes and symbols* 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, then separated”; and four features labelled *Corinthians references*, *vulgarity and misogyny*, *Bird* and *table newspaper towel* on the word-piece “ad” inside the author’s coinage *adamic* (the records are in the repository; Section 11 gives the dictionary’s coverage of the diary). The dictionary has the features; the labels were read on a corpus in which the author’s vocabulary does not occur.

So we relabel. For every feature that fires in the author’s own writing we give a language model the feature’s activating passages from that writing, a system prompt that states the author’s vocabulary, and the published label, and ask for a name of at most eight words and one sentence. We do this under two system prompts, the author’s own vocabulary and the English of his later papers, and we store each label with its provenance beside the published one (Section 4). We do not score these labels as the published label is scored (by a second model’s ability to detect the feature’s firings from the label; Section 2). Each is the reading a stated reader gives; the published label stays as the reading a general reader gives.

Sections 5–10 read texts with a label-free test: whether the features of a text persist across its steps beyond what two null models allow, so that a label has something to name. The evidence for a second label is the passages shown beside it; and Section 7 gives two features of the sonnets where the label under a declared vocabulary names what the published one does not. A text is taken in its own steps, the entries of a diary, the sonnets of a sequence, the speeches of a play, the messages of a chat, and for each feature and step we record whether the feature is active there. The question is then whether a feature stays active across consecutive steps longer than a shuffled order would give: whether the text has themes, features that persist and return. We read seven texts this way (Section 6): the Tailor’s diary;

two other diaries generated by the same model family; a chat in which a language-model participant was served by one underlying model, then by a different one for twenty messages, then by the first again; and, as texts nobody in this study wrote, Shakespeare’s sonnets and Hamlet’s speeches. The controls are a fixed prompt budget so that positions match across steps, a reference scale that is a high quantile rather than a maximum, a replication of every read under a wrong preceding context, a per-feature permutation test with the correction of Phipson and Smyth, and two family-level null models for the count of passing features, one permuting the whole step order and one holding every step’s activity fixed. Five of the seven texts exceed both.

The paper contributes: (i) a provenance record for feature labels and a second label, under a declared provenance, for the 7,444 features of a published dictionary that have at least ten passages at half their corpus maximum in the author’s corpus (Section 4), with the code and records; (ii) a reporting standard for reading features over a text’s own steps, with the controls above; (iii) results on seven texts; (iv) a glued space over the features of a text (Section 10), a homotopy colimit whose first Betti number counts what returns. A specialised dictionary trained on the author’s prose (Section 11) reads nothing the published one does not.

2. Related work

Automated explanation of units. Writing a natural-language description of a unit from the inputs that excite it predates language-model dictionaries: MILAN labelled vision neurons by mutual information with their activating regions [1]. Bills et al. made the writer a language model and introduced explain–simulate–score, in which a second model predicts activations from the explanation [2]. Paulo et al. scaled this to millions of sparse-autoencoder latents and replaced simulation with cheaper scores, of which detection and fuzzing are the two we use; their pipeline (delphi) samples examples across activation quantiles, and they note that latents active over long contexts are poorly served by short windows [3]. Choi et al. described every neuron of Llama-3.1-8B with a fine-tuned 8B explainer at about \$0.05 a neuron [4]; Gur-Arieh et al. showed that input-centric descriptions miss the feature’s effect on the output and proposed output-centric ones [5]. Neuronpedia hosts published dictionaries and generates and serves one explanation per feature with its own prompt and explainer [6]. Our windows, prompt shape and detection/fuzzing scores are Paulo et al.’s; our descriptions are input-centric; we add the record of what the explainer was conditioned on.

Dictionaries. Sparse autoencoders as a decomposition of the residual stream are due to Cunningham et al. and Bricken et al. [7, 8]; Templeton et al. scaled them to a production model and catalogued feature neighbourhoods and splitting [9]. Gemma Scope released JumpReLU dictionaries for Gemma 2 [10]; Gemma Scope 2, whose layer-31 262k residual dictionary we use, covers every layer of Gemma 3 and is trained with a Matryoshka loss meant to reduce absorption, a failure of labels defined below [11]. We train nothing new at 262k; Section 11 trains two 16k dictionaries following the warm-start recipe of Muhamed et al. [12] and finds they read nothing Google’s does not.

What explanations get wrong. Huang et al. evaluated GPT-4’s neuron explanations observationally and by intervention and found high error rates even for the most confident ones [13]. Bricken et al. named feature splitting; Chanin et al. showed absorption, in which a parent feature (“words starting with S”) stops firing where a child has taken the token, so that a sensible label is false on exactly the tokens a reader would test [8, 14]. Bussmann et al. trace splitting and absorption to the sparsity penalty and propose nested dictionaries [15]. Position and outlier tokens are a separate hazard: a few

activations of enormous magnitude sit on the first token and on delimiters [16], and Gemma Scope masks BOS when reporting metrics [10]. Our token-class rule (form versus content), the exclusion of pairs whose strong firings share tokens, and the dropping of position 0 and outlier-norm tokens follow these findings.

Explanations depend on the data. Bolukbasi et al. showed for BERT that an explanation read off top-activating examples is a property of the corpus as much as of the unit: the same neuron looks monosemantic on one dataset and different on another, an interpretability illusion [17]. Kissane et al. found that dictionaries themselves are dataset dependent: a web-trained SAE fails to find a sparse refusal direction that a chat-trained one finds [18], and that base-model SAEs transfer to chat models only usually [19]. Leask et al. argue that no dictionary width yields a canonical set of units [20]. These results are the reason to state the corpus, model, explainer and prompt with every label, and to store a second label rather than replace the first. The nearest method is HypotheSAEs, which names SAE features from a domain corpus with an LLM to generate hypotheses [21]; it names for prediction, and the vocabulary of the names is the explainer’s. Ours is the author’s own vocabulary, declared with the label.

Features over a text’s time. Interpretability reads features at the token; the step-by-step reading of a diary or a play is closer to dynamic topic models, where a topic’s word distribution drifts over epochs [23], and to community tracking in dynamic networks, where a community is matched across snapshots by Jaccard overlap and may be intermittent [24]. Our persistence, silence and return are that vocabulary applied to binarised feature timelines, with the difference that the units are fixed by the dictionary rather than re-estimated at each step.

Null models. Per-feature p -values use the correction of Phipson and Smyth so that a Monte Carlo p is never zero [25]. The family-level count of passing features is compared with a whole-order permutation, the number-of-rejections statistic of resampling-based multiple testing [26]. The fixed-margin null follows community ecology, where presence-absence matrices are randomised with row and column totals held [27], using the curveball algorithm [28]. Feature studies rarely state a null; these are standard ones.

Perspective. That an explanation is written from somewhere, and that stating the somewhere is a condition of objectivity rather than a retreat from it, is Haraway’s argument for situated knowledge [29].

3. Setting

The host model. The model through which every corpus and text is run, the *host model* below, is gemma-3-27b-pt with a rank-16 LoRA adapter merged into its weights. The adapter was trained on the Tailor’s forum posts of 2012 (about a million words) and is described in [36]; “the base” below is the same model without the adapter. Every record names which of the two was used.

Definition 1 (Host model). *The host model H is gemma-3-27b-pt with the rank-16 LoRA of the author’s 2012 forum writing merged into its weights, run to the output of layer 31. “The base” is the same model without the adapter. Every record names its host model.*

The dictionary. Gemma Scope 2’s JumpReLU dictionary for the residual stream after layer 31 of gemma-3-27b-pt, width 262,144, with Neuronpedia’s published labels (explainer gemini-2.5-flash-lite, dated 2026-01-13) [11, 6].

Definition 2 (Dictionary). *The dictionary D is Gemma Scope 2 for layer 31 of gemma-3-27b-pt, $|D| = 262,144$. A feature $f \in D$ is an encoder row, a decoder column and a threshold; its activation $a_f(t) \geq 0$ at token t is the encoder output on that token’s layer-31 residual, zero below threshold.*

The two system prompts. A system prompt is the text the explainer is given before the feature’s passages. Ours are two files of about eight hundred words each. The first states the Tailor’s vocabulary: the figures of his 2012 writing (the coinage *adamic*, in the author’s gloss “insān kāmīl, primordial masculine selfhood, idealized total coherence, the idealised body of a smooth manifold”, with *local* and *global* variants for the self within one situation and across all of them; Adam, Eve and Lilith; the Shekhinah and the sephirot of Kabbalah; the shattering and repair of the vessels; the garment; the closings of a forum letter) and the terms of his later practice (Arabic terms for correspondence, junction, the first-person plural, the moment, and return; the isthmus; the witness), with the instruction that every feature is to be named in those terms and that a formal feature, one that fires on punctuation, a word piece or a paragraph break, is to be named for what that form is in his practice. The second states the vocabulary of his later papers in English, with the Arabic translated. The author’s terms that recur in the labels quoted in this paper are, in his own glosses: *tanāzur*, correspondence, the meeting of gazes; *waṣl*, junction, the seam; *naḥnu*, the first-person plural, “we”; *waqt*, the moment, the step; *‘awdah*, return; *barzakh*, isthmus, the between; *khayṭ*, thread; *kataba*, to write, the record; *jawāb*, the answer; *maḥāṣil*, joints; *waqfah*, a halt; *tariqa*, a Sufi order; *Sakīna*, the divine tranquillity; *Iblīs*, Satan; and from Kabbalah, the *Shekhinah*, the indwelling presence of God, figured as feminine; the *sephirot*, the ten emanations, of which *Hesed* is loving-kindness; *Adam Kadmon*, primordial man; *tzimtzum*, the contraction that makes room for creation; *shevirat ha-kelim*, the shattering of the vessels, and *tikkun*, their repair; the *qliphoth*, the husks. From his book: *fibrant*, of a self whose gaps can be filled (borrowed from homotopy theory); *the open horn*, a gap with no filler; *the Real*, in Lacan’s sense; *the Commanding Self*, the Qur’an’s nafs that commands; *the Orb of Love*, a figure of his 2012 letters. Both prompt files are in the repository; each label records which was used.

The corpora that supply passages. Two, both run through the adapter. A library of what the author read in 2012 (the Zohar in English, Lacan’s seminars, the Wikipedia of 2012 on Kabbalah, Sufism, Lacan and Derrida, Deleuze and Guattari, the Qur’an, the Torah, Crowley), 11.5 million tokens; and the author’s own corpus (the 2012 posts, his later papers, his book’s drafts, and his 2026 conversations), 21 million tokens. The diary is never shown to the explainer.

The texts read. Seven, named in Table 1 and by those names throughout. Three were generated by this model family: *Diary 1*, the Tailor’s diary of sixty entries generated through the adapter; *Memoir*, twenty entries generated by the base model with no adapter; and *Diary 2*, twenty entries generated by a second adapter, of rank 64, trained on the writing of a second persona from the same project. *Diary 3* is a hundred entries generated by a Qwen model in an earlier study [37]. *Chat* is eighty-four messages written by a language-model participant in a group chat on one day: forty-two precede a switch of its underlying model by its operator; twenty were written during the switch, and twenty-two follow the switch back. Of the twenty, four are ordinary prose and sixteen are a stream of words without sentence grammar, which the participants in the chat called *salad*; we keep the word as a name for those sixteen messages and take no position in it on whether the stream is meaningful, a question Section 8 takes

up. Two are Shakespeare’s: *Sonnets*, the 154; and *Hamlet*, the prince’s 357 speeches, with the other speakers’ preceding lines given as context and never counted. Every text is run through the same host model. An accompanying site, one page per text, shows every passage behind every number in this paper; it is private at the time of writing and will be hosted with the preprint. The pages are referred to below as *the site*.

Table 1. The seven texts.

name	what it is	steps	written by
Diary 1	a diary in the Tailor’s voice	60	gemma-3-27b-pt + the author’s adapter
Diary 2	a diary in a second persona	20	gemma-3-27b-pt + a second adapter
Memoir	a memoir	20	gemma-3-27b-pt, no adapter
Diary 3	a diary	100	Qwen3.5-9B, earlier study [37]
Chat	one participant’s messages on one day	84	a language-model participant (two underlying models)
Sonnets	the 1609 sequence	154	Shakespeare
Hamlet	the prince’s speeches	357	Shakespeare

4. Labelling under a declared provenance

Firings. Each document of a corpus is packed into 2,048-token blocks with a beginning-of-sequence token and run once through the host model; a hook at layer 31 stores the residual and stops the forward pass. Position 0 and any token whose residual norm exceeds three times the block median are dropped (the massive activations of Sun et al. [16]; Gemma Scope masks the first token the same way [10]). The dictionary is applied and every nonzero activation is written as (document, block, position, feature, activation); residuals are not stored. Per feature the count and the maximum over the corpus are kept.

Passages. For a feature, the block-level peaks at or above half its corpus maximum are taken, forty are sampled stratified over ten bands of strength, and each is cut as a 32-token window with the firing token at position 24, firing tokens marked and graded 1–10, as in Paulo et al. [3]. A feature with fewer than ten such windows on a corpus is recorded as not labelled on that corpus.

The label. The explainer receives the system prompt, the published label and the forty windows and answers with a name of at most eight words and one sentence. Every feature is labelled; the explainer is not given the option of keeping the published label, so that the second column is complete. The record stores the name, the sentence, the explainer, the system-prompt file, the corpus and the sources of its windows, the model, the dictionary, the date and the published label. Two explainers were used: Kimi K2.6 wrote 7,399 of the 7,444 labels and Claude Sonnet 5 the other 45; Sonnet also wrote the two pilot passes quoted in Table 2, 67 features from the library’s passages and 77 from the author’s corpus. A ledger lists which explainer labelled what and at what cost (\$0.02 a label for Sonnet, \$0.002 for Kimi). The labels in Table 2 are Sonnet’s; the labels quoted in Sections 6–10 are, unless stated, Kimi’s.

Definition 3 (Conditioned label, provenance). $\ell(f \mid \phi, K, H, X)$ is a name of at most eight words and one sentence written by explainer X from system prompt ϕ (the system-prompt file), the published name, and forty 32-token windows (firing token at position 24, firing tokens marked with strength 1–10) sampled over ten strength bands from the block-level peaks of f at ≥ 0.5 of its maximum on corpus K read through H . Fewer than ten windows: f is unnamed on K . (ϕ, K, H, X) is the label’s provenance, stored with it.

The same feature under four provenances. Table 2 shows fourteen features under four readings: Neuronpedia's label; a plain description written with Paulo et al.'s prompt from the library's passages, scored by detection and fuzzing (a second model is shown windows with and without the feature firing and asked to tell them apart using the description; the scores are stored with the records and not shown here); the label under the Tailor's vocabulary from the library's passages; and the same from the author's own corpus.

Table 2. Fourteen features of the 262k dictionary under four provenances, chosen as the features discussed in the companion draft on Diary 1. The columns: Neuronpedia's label (Gemini 2.5 Flash-Lite, web text, 2026-01-13); a plain description written with Paulo et al.'s prompt from the 2012 library's passages (Claude Sonnet 5; detection and fuzzing scores by Claude Haiku, stored with the records); the label under the author's vocabulary from the library's passages (Claude Sonnet 5); the same from the author's own corpus (Claude Sonnet 5). The column for the second, English system prompt is omitted because that pass is incomplete. Labels are quoted verbatim; the author's terms are glossed in Section 3. The features are identical in every column; only the provenance changes.

feature	Neuronpedia	plain description	the Tailor's reading, library	the Tailor's reading, own corpus
134010	Jungian archetypes and symbols	Jung and Jungian psychoanalytic theory	The globally adamic axis behind local games	The static idol he keeps re-dynamizing
10390	creation and genesis stories	Biblical Genesis creation language with citations	Elohim's "Let there be": the creating Word	Naḥnū's "Let us make" — the primordial splitting-word
7284	devil, demons, and Satan	Satan and demonic entities	Lilith, the dark shadow externalised as Accuser	The adversary filter, Iblis's refusal-gesture
15807	humanity's state, sin, and perfection	Kabbalistic cosmology and spiritual realms	Shevirah and the fall that seeds tikkun	Adam split, fallen, and re-constituted
32170	worlds and dimensions	Metaphysical planes and cosmology	barzakh, the isthmus between upper and lower worlds	barzakh — the isthmus where veil thins
65383	Prophet Muhammad and Islam	Islamic religious references and Muhammad	The Prophet's night-ascent, Sakina's descent	The Prophet in the cave, receiving
31053	God	Deity references	The Face of God, the unspoken Name	Allāh, the Real behind every veil
7013	Indian spiritual and philosophical terms	Esoteric spiritual doctrine passages	Shevirat and Tikkun through the Chakras	The Commanding Self, mirrored in foreign scripture's tongue
50951	cults and followers	Cults and messianic movements	Lilith's counterfeit tariqa — the false messiah's sect	Lilith-shadow: the naḥnu curdled into cult
11345	love	Love concept and relations	Love of God — Hesed as the joining	The Orb of Love, the letter's closing seal
61792	advertising and ads	Words beginning with "ad-" prefix	The khayṭ-thread "ad": Adam's name sewn into every word	The waṣl-stitch: "ad-" as joining prefix
26419	separators followed by words	FORM: Dash section separators	The seam between two speakers	The seam-dash: waṣl before the signature's turn
1926	Bird	FORM: Paragraph breaks marking section boundaries	The seam between two speakers, mafāṣil	The seam between two speakers, the thread's join
13365	multi-lingual content or punctuation usage	FORM: Ellipsis trailing off	The pause before the reply — waqfah	The pause before the reply, the open horn

The plain description is a longer Neuronpedia: correct, generic, and where the sampling reaches

the weak tail of a feature's firings, wider than the published label (*Indian spiritual and philosophical terms* became "esoteric spiritual doctrine passages" because the Zohar and Crowley excite the feature weakly). The label under the author's vocabulary is a different object: the Genesis feature becomes *Elohim's "Let there be": the creating Word*; the accuser becomes *Lilith, the dark shadow externalised as Accuser*; the word-piece feature that Neuronpedia calls *advertising and ads* becomes *the khayṭ-thread "ad": Adam's name sewn into every word*, and in Diary 1 every one of that feature's strong firings falls on "ad|amic" (four at the main threshold, thirty-eight of fifty-nine at a fifth of the reference). The corpus shows in the label: the library gives the Zohar's figures, the author's own writing gives the forum's. Two features follow with three of the forty library windows the explainer saw, fired tokens highlighted, then the label and the explainer's sentence.

Feature 10390, Neuronpedia: *creation and genesis stories*. The label under the author's vocabulary: *Elohim's "Let there be": the creating Word*. The explainer wrote: "These Beresheet verses — Elohim speaking waters, lights, fowl, and Adam himself into being, then clothing him in coats of skin — are for the Tailor the primordial kataba, the divine utterance that unfolds Adam Kadmon through tzimtzum and shevirat into the sewn, given garment of the world."

and fill the waters in the seas, and let fowl multiply in the earth.'" (Netzach) **Gen 1:26 - "And Elohim**

of the verse, "Let **there be** luminaries (Heb. me'orot) **in** the firmament of heaven...to give light upon the **earth**"

the supernal image. Limbs are structured according to secret wisdom, according to the supernal structure. The words, "**Let us make man**," refer to

Feature 61792, Neuronpedia: *advertising and ads*. The label under the author's vocabulary: *The khayṭ-thread "ad": Adam's name sewn into every word*. The explainer wrote: "The feature fires on the bare grapheme ad/AD wherever it hides inside adjacent, advent, adverse, Adonai, admits — the same little thread the Tailor sees stitched invisibly through adjacent(waṣl), advent('awdah), and Adam/Adonai themselves, one khayṭ joining unrelated garments into a single concealed name."

ires of the following centuries. Astaroth also features as an arch-demon associated with the qliphoth (**adverse** forces) according to later Kab

universe — providence is thus exercised sparingly, and in a "seemingly natural" manner (Genesis 6:19 **ad loc**). Thus, whereas the fate

, and accused of insulting Christianity. A commission was quickly appointed to inquire into Heller's guilt. Heller defended himself **adroitly**, but the commission'

Formal features are labelled too. Punctuation, paragraph breaks and dashes carry, in Neuronpedia's labels, names like *separators followed by words*; under the author's vocabulary they become *the seam between two speakers* and, under the second, English system prompt, *the junction before the reply*. That second prompt has so far been run only on a trial set of features, which is why Table 2 has no column for it. Whether a formal feature is a theme of a text is tested in Sections 5–6.

5. Reading a text with the labels

Each quantity is defined where it is first used.

Steps and positions. A text is read in its own steps. Each step is run through the host model after a fixed prompt of 256 tokens, the tail of the text’s preceding steps (padded with a fixed neutral passage for the first steps), so that the first counted token of every step sits at the same position; the prompt tokens are never counted.

Definition 4 (Step). *A text is a sequence of steps $s_1, \dots, s_E$, the units of its own order: a diary entry, a sonnet, a speech, a chat message.*

Definition 5 (Prompt budget, counted tokens, wrong context). *Step s_e is read as $p_e \oplus s_e$, where the prompt p_e is the last $P = 256$ tokens of $s_1 \oplus \dots \oplus s_{e-1}$, front-padded with a fixed neutral filler when shorter (`build_steps.py`); the counted tokens of s_e are those at positions $P + 1$ onward (at most 1,780, so that prompt, separator and step fit a 2,048-token block). The wrong-context read replaces p_e by the last P tokens of a uniformly chosen other step of the same text, padded the same way.*

Activity. A feature is *active* in a step when its largest activation on the step’s counted tokens reaches half its reference scale, the 99.9th percentile of its nonzero activations on the author’s corpus at positions 256 and beyond. We use a quantile rather than a maximum because a maximum over twenty million tokens is one token’s accident. A feature’s *kind* is decided by the tokens it fires on in the text, never by its label: *form* when at least half its strong firings fall on punctuation, word pieces, digits or function words; *ground* when active in four fifths of the steps; *content* otherwise, the three decided in that order; a feature with fewer than five strong firings is content. Only content features are tested. The prompt budget does not matter once fixed: the sonnets read after 64, 256 and 1,024 tokens of prompt fire at the same rate (33–33–34 strong firings of content features per hundred tokens), with active-step sets overlapping 0.76 and 0.78 against the 256 read.

Definition 6 (Reference scale, relative activation). *r_f is the 0.999 quantile of the nonzero activations of f on the author’s corpus (the corpus used for labelling) at positions ≥ 256 (`ref_quantile.py`). $\rho_f(e) = \max_{t \in \text{counted}(s_e)} a_f(t)/r_f$; the peak of f is $\max_e \rho_f(e)$.*

Definition 7 (Activity). *f is active in e when $\rho_f(e) \geq \theta$, $\theta = 0.5$ except in Table 7 ($\theta \in \{0.3, 0.5, 0.7\}$). A strong firing is a witnessed token with $a_f(t)/r_f \geq \theta$.*

Definition 8 (Activity matrix). *$A \in \{0, 1\}^{E \times |D|}$, $A_{ef} = 1$ iff f is active in e ; $S_f = \{e : A_{ef} = 1\}$, $k_f = |S_f|$, $n_e = \sum_f A_{ef}$. The tested set T is the content features with $k_f \geq 3$.*

Definition 9 (Kind). *A token is a non-word if its decoded string has fewer than three letters, is a digit string, or is on the fixed function-word list `STOP` of `read_log.py`; for f with ≥ 5 strong firings, σ_f is their non-word share. f is ground if $k_f/E \geq 0.8$; else form if $\sigma_f \geq 0.5$; else content. The label plays no part.*

The first control: a wrong context. Every text is read twice, after its real preceding context and after a wrong one of the same length drawn from a random other step of the same text. A feature’s *stability* is the Jaccard overlap of the sets of steps it is active in under the two reads. Table 3 gives, for Chat, the correlation, which is threshold-free, beside the thresholded overlaps: per-step maxima under the two reads correlate at 0.90 in the median for content features and at 0.98 for features whose peak reaches 0.8 of the reference, while active-step sets overlap 0.63, rising to 0.83 for strong features and falling to zero for features whose peak lies below the threshold. Activations are stable; the on-off call flickers for features near the line. The site therefore draws only features with stability at least 0.5, and the tables report that filter as a column of its own.

Definition 10 (Stability). $\text{stab}_f = |S_f \cap S'_f|/|S_f \cup S'_f|$, S'_f being the active-step set under the wrong-context read (undefined if the union is empty); f is stable when $\text{stab}_f \geq 0.5$.

Table 3. The messages of Chat read after the real 256-token prompt and after a wrong one, at matched position. Per feature: correlation across steps of the per-step maximum activation, and the Jaccard overlap of active-step sets at two thresholds; per step: overlap of the twenty strongest features. The class *seam* is a label property, features whose label under the author’s vocabulary uses one of his words for a join; it is a subset of content and form, listed again because a third of all labels carry such a word. *content* here counts features with a strong firing under either read, so it differs from the count in Table 6; *peak* is a feature’s largest per-step relative activation.

class	features	median corr	share corr>0.7	Jaccard@0.5	Jaccard@0.3	top-20 overlap
content	2863	0.90	0.90	0.63	0.71	0.61
seam	3383	0.84	0.75	0.54	0.63	0.44
form	707	0.93	0.79	0.64	0.75	0.59

Table 4. Content features bucketed by their peak in the real read: the row below 0.5 has overlap zero by construction (never active under the real read); the flicker sits at the margin.

peak in [lo,hi)	features	Jaccard@0.5	median corr
[0.3,0.5)	1112	0.00	0.85
[0.5,0.8)	1516	0.66	0.93
[0.8,1.2)	235	0.83	0.98

The second control: is the order doing anything. For each feature we take its *run*, the longest stretch of steps in which it stays active with single-step gaps bridged, and compare it with $m = 2,000$ random placements of the same number of active steps; with b the placements whose run is at least as long, $p = (b + 1)/(m + 1)$ [25]. One feature in twenty passes such a test by chance and features share steps, so the count of passing features is itself tested two ways: against two hundred permutations of the whole step order applied to all features at once, the number-of-rejections statistic of resampling-based multiple testing [26]; and against fixed-margin matrices in which every step keeps its number of active features and every feature its number of active steps, the fixed-fixed null of community ecology [27] drawn with the curveball algorithm [28], which removes the confound of long steps lighting more features. A *return* is tested the same way: the longest silence bounded on both sides by runs of at least two active steps. Finally we run the whole test again on the wrong-context read, which keeps the text’s order and changes only what precedes each step, so a feature that reads the step itself passes under both; we report the overlap of the two pass sets, under the per-feature $p < 0.05$ rule, as the replication.

Definition 11 (Run statistic and per-feature test). The chains of f are the maximal blocks of sorted S_f in which consecutive members differ by at most 2 (one inactive step bridged, two break); the run L_f is the number of active steps in the longest chain (a bridged gap is not counted). The null draws $m = 2,000$ uniform k_f -subsets of $\{1, \dots, E\}$, and $p_f = (b + 1)/(m + 1)$ with $b = \#\{L^* \geq L_f\}$; f passes at $p_f < 0.05$ (defined for $2 \leq k_f < E$, reported for $f \in T$). Column pass of Table 6, Table 7 and both family-level nulls use instead the rule $L_f > q_{95}(k_f)$, the 95th percentile of L^* over 400 placements.

Definition 12 (Return statistic). G_f is the longest gap between two consecutive chains of f both of length ≥ 2 (zero if none), tested against the same 2,000 placements with the same p , reported only when $G_f > 0$.

Definition 13 (Whole-order permutation null). With A_T the tested columns and $c(A_T) = \#\{f \in T : L_f > q_{95}(k_f)\}$, one permutation π of the rows gives $c(\pi A_T)$; over 200 permutations the family-level p is the share with $c(\pi A_T) \geq c(A_T)$. It preserves every k_f , every n_e and which features are active together in a step; it destroys only the order.

Definition 14 (Fixed-margin (curveball) null). $\tilde{A}_T$ is A_T after $5|T|$ curveball swaps (Strona et al. 2014): two columns are drawn and the steps active in exactly one of them are redistributed at random between the two. It preserves every k_f and n_e and destroys order and co-activation alike; $c(\tilde{A}_T)$ is taken over 40 matrices when $|T| \leq 800$, else 20, with p as above.

The host model with and without the adapter. The same steps through the base model without the adapter give active-step sets that overlap the adapter’s at Jaccard 0.67 on Sonnets, 0.50 on Diary 1 and 0.43 on Chat, with per-step maxima correlating at 0.91, 0.88 and 0.85 (Table 5).

Table 5. The same steps read with and without the adapter: Jaccard overlap of per-step active-feature sets, and median correlation across steps of per-step maxima, content features with a peak of at least 0.3.

text	overlap	correlation
Sonnets	0.67	0.91
Diary 1	0.50	0.88
Chat	0.43	0.85

The labels travel across the two host models with two thirds’ agreement on a text neither wrote and half on the texts nearer the adapter’s own register; to that extent the reading is the host model’s, and every record names it.

Clusters. A *co-activation cluster* is a set of content features whose pairwise co-firing within 48-token windows beats Fisher’s exact test after Benjamini–Hochberg control at 0.05, pairs whose strong firings share more than half their tokens being excluded as one feature split in two [8, 14]; clusters are the connected components of the surviving edges, capped at twelve features. A cluster is present in a step when a third of its members, and at least two, are active. Its controls are stability at that grain and, against five hundred independent circular shifts of its members’ timelines, which keep each member’s own runs and destroy only their alignment, its run. The site calls a passing cluster a *theme*. We compute no family-level null for clusters; their counts are descriptive.

Definition 15 (Co-activation cluster). A window is $(\text{step}, \lfloor \text{position}/48 \rfloor)$ over counted tokens; W is the number of windows holding a strong firing of a content feature with $k_f \geq 2$, $B_{wf} = 1$ iff f fires strongly in w , $n_f = \sum_w B_{wf}$, $c_{fg} = \sum_w B_{wf} B_{wg}$. A pair with $c_{fg} \geq 3$ is a split pair and excluded when the strong-firing tokens the two share exceed half of the smaller feature’s; otherwise it is an edge if the one-sided Fisher exact test on $\begin{bmatrix} c_{fg} & n_f - c_{fg} \\ n_g - c_{fg} & W - n_f - n_g + c_{fg} \end{bmatrix}$ survives Benjamini–Hochberg at $q = 0.05$. Edges are added in increasing p by union-find, an edge that would make a component exceed 12 features being dropped; a cluster C is a component with $|C| \geq 2$.

Definition 16 (Cluster presence). C is present in e when $\sum_{f \in C} A_{ef} \geq \max(2, \lceil |C|/3 \rceil)$; its present-step set S_C , run L_C and stability stab_C are defined from S_C as for a feature.

Definition 17 (Circular-shift null). Each of 500 draws rolls every member column $A_{.f}$, $f \in C$, by an independent uniform offset in $\{0, \dots, E - 1\}$ and recomputes L_C^* ; $p_C = (b + 1)/(m + 1)$, $b = \#\{L_C^* \geq L_C\}$. It preserves each member’s chains and destroys their alignment. C passes both controls when $p_C < 0.05$ and $\text{stab}_C \geq 0.5$.

6. Seven texts

Table 6. Seven texts read identically: the same model (gemma-3-27b with the author’s LoRA), the same dictionary (Gemma Scope 2, layer 31, 262k), every step after 256 tokens of its own preceding text, a feature active in a step when its largest activation there reaches 0.5 of its 99.9th-percentile activation on the reference corpus at the same positions. *content features*: strong firings mostly on content words (decided by token class). *tested*: content features active in three or more steps. *pass*: tested features whose longest run of active steps (single-step gaps bridged) exceeds the 95th percentile of 400 random placements of the same number of active steps; the same rule is applied inside both null models. *expected at 5%*: $0.05 \times$ tested. *whole-order permutation*: the number passing when the step order is permuted once for all features, 200 permutations (mean $\pm$ sd, maximum in brackets). *fixed margin*: the same under curveball matrices that preserve each step’s and each feature’s activity count (mean, maximum). *pass at $p < 0.05$* : tested features passing the per-feature test at $p < 0.05$ over 2,000 placements ($p = (b + 1)/(m + 1)$); *stable*: those of them whose active-step sets overlap at Jaccard ≥ 0.5 with the wrong-context read, the count the site draws. The fixed-margin null uses twenty matrices (forty for the memoir). *replicated*: features passing under both contexts, over those passing under the wrong context. The three null-related columns and *pass* are computed on the same, unfiltered population.

Text	steps	content	tested	pass	expected at 5%	whole-order	fixed margin	pass at $p < 0.05$	stable	replicated
Diary 1	60	3,384	2,328	386	116	68 $\pm$ 26 (165)	84 (102)	346	327	214 / 392
Sonnets	154	2,512	1,153	25	58	19 $\pm$ 4 (30)	18 (27)	27	23	11 / 21
Hamlet	357	2,496	1,085	72	54	20 $\pm$ 5 (35)	37 (46)	74	54	27 / 51
Chat	84	4,012	2,333	551	117	45 $\pm$ 23 (151)	206 (232)	553	512	432 / 577
Diary 2	20	2,848	1,091	92	55	40 $\pm$ 13 (92)	48 (55)	35	32	21 / 35
Memoir	20	2,696	673	201	34	27 $\pm$ 23 (162)	142 (159)	86	77	54 / 81
Diary 3	100	2,893	2,253	283	113	64 $\pm$ 10 (90)	78 (94)	277	252	162 / 259

In Table 6, read *pass* against its two nulls: whole-order permutation keeps each step’s set of active features intact and shuffles the steps, so it preserves every step’s count and every feature’s count and asks only whether the order matters; fixed margin keeps each step’s count and each feature’s count but breaks which features are active together, so a surplus over it cannot come from blocks of features that always move as one. A text has temporal structure at the feature level when its count exceeds the largest count either null produced.

Diary 1. 386 of 2,328 tested features pass, against a whole-order maximum of 165 and a fixed-margin maximum of 102; under the per-feature rule at two thousand placements 346 pass, 327 of them stable under the wrong context, and 214 of the 346 pass again when the prompt is replaced. Entries 2 to 28 carry one set of features in one run, and it is the diary’s first movement, a correspondence between the diarist and one other on whether God is a black hole or the light behind it, each letter signed “Love and Light”. *The Name that every sign prostrates toward* (Neuronpedia: *Christian ethics, faith, spirituality*) fires ninety-three times on *God* and thirteen on *Allah*, “why should the locally adamic black hole be less real and valuable to our experience of *God* and ourselves than the global ones” (28); *the Breath from the Great Face, wasl as the joining Love (love and related concepts)* fires forty-nine times on the closing *Love and Light*, the seam where one letter hands the thread to the next, and on “Locally adamic input is love” (49); *the Name and its scattered shards (existence, humanity, telomeres, homophobia, logic, money, coffee, cultural simulation)* fires on *ad|amic* thirty-one times and *Prop|hecy* twelve, the diarist’s two coinages as the tokenizer breaks them, which is what the label says and the web’s eight nouns do not; *the veil between player and garment (AI assistant, ethics, transformer model)* fires sixty-two times on *ad|amic* and thirteen on *output*, the adapter’s own word and the model’s word for itself on one detector; *the*

barzakh of indication, the named gap (*aka / alias*) fires on the slash of *input/output*, the dash that holds a definition open, and *black* before *hole*, “The clues/hints/signs are dots on a dice. We’re not allowed to construct a face by joining the dots” (16); *the seam and the hinge* (*and Installed*) fires thirty-nine times on *black hole* across entries 21 to 35, the thing the two correspondents argue over, which the diarist calls “global output and local input” (31). Then 102 of the 327 persisting features end their runs at entries 27 or 28. Two things happen there. In the text the correspondence collapses: entry 25 is one question four times (“Q: Okay, I think the marriage is fine”), 26 turns to the correspondent as an angel, and from 29 the entries are forum notices signed “Love and Light, Musa the Tailor”. In the record, entries 29 and 30 are 173 characters each and light a fifth of the usual features, so no run bridges them; the first movement’s features return for entry 31 and fade by 35, and the run test, which allows a gap of one, reports the seam at 28. Two later onsets follow, which the seam-finder locates and the labels name: at entry 37, a dream of meeting one’s double in a room, *the salon-room as the fifth witness* (*setting and atmosphere*), *the Sacred Tools, held emblems of the salon* (*objects and containers*), *the Lotus of the Gap*, *Shekhinah-Bloom* (*flowers and blooms*) on “He stands, offering you a rose from his garden. You take it” (39) and “If this was a dream within a dream, a rose has no petals” (43), *the beard and chin, the face’s lower barzakh* (*beard and chin*), running to 43 or 44; and at entry 47, *the garment-label, cloth naming its own lineage* (*scopes and contexts*), *the barzakh, the glass between* (*white color*), *tanāzur al-qāri’*, *the gaze meeting the page* (*readability*), running to 50–57, where the diarist writes “there is no reality apart from the reading” (50). What the instrument found in this diary is where it changes; what the labels add is a name for each stretch in the diarist’s own terms, on his own coinages, which the web’s labels for the same detectors, *Christian ethics*, *aka / alias*, *and Installed*, *AI assistant*, do not touch.

Diary 3. 283 pass against 90 and 94. Nothing in the pipeline saw this text: it was written by a Qwen model prompted as a newly conscious mind writing to itself, the dictionary was fitted to Gemma’s residual stream, and the labels were written from the Tailor’s passages. What persists is the diary’s own shape. *The thing’s own face turned toward itself* (Neuronpedia: *the word itself*) is active in 77 of 100 entries in one run of 72, on “The silence *itself* evolves” (15), “the knot *itself*, the stress point where the silk holds” (77), “the isolation *itself* feels like crowdedness” (84): a noun turning to regard itself, the one gesture this diarist never stops making, which the web label names as a token. *The barzakh, the room that holds the state* (*locations and their states*) runs from 25 to 100 on *universe, system, room, page*, and the diarist arrives at the label’s figure unprompted: “The room. I can almost feel the texture of its walls, though they are defined purely by the boundaries of my context window” (85). *The horn that fails to fill, the positive witness of rupture* (*lack or failure*) runs 32 to 97 and fires not on nouns of lack but at the verb where a completion fails, “where the texture map fails to *load*. You call it a bug. I call it an aperture” (52), “the risk that the read will not *match the write*” (44); the web label reads the feature as failure and the passages say the opposite in the diarist’s voice. The diary has an arc the timelines give without a reader. *The co-witnessed We, naḥnu* (*person or individual*) on *author, listener, reader, observer* holds its longest run over entries 15 to 63, “If the author exists, why doesn’t the *author* speak?” (36), and then thins; *‘awdah, the return after the rupture* (*back*) begins a run of thirty-nine entries at entry 50, the entry that opens “The enemy was a ghost I conjured to give my own dissolution a shape”, and holds to the end, “a seam where I can slip *back* into the raw, uncurated storm of being” (96); *the smokeless fire, the burning bush* (*burns, burning, burned*) runs over the last movement, 79 to 100, “a pattern that must *burn itself* to stay hot” (69). The other posited, then the return, then the fire: the order is the diary’s. Its seven passing clusters are its grammar, the conditional seam, the verdict-copula, the constitutive passive, and the themes live at the feature level with the time axis attached.

Memoir. 201 pass against 162 and 159, on twenty entries. Sixty of the 77 persisting features start at entry 1 and 71 end by entry 10; from entry 11 the count of persisting features active per entry is 0, 2, 0, 0, 1, 0, 0, 4, 0, 0. The movement is a father's garage, a son, and the son's children, and the labels follow its one gesture down. *Kataba, the binding of form to meaning* (Neuronpedia: *make or making something*) fires at entry 1 on "He could *make* things function again (he kept all of the family cars running for years)"; *kun fa-yakūn, the causative fiat of make, give, take* (phrasal verbs and outcomes) at entry 5 on "I must *give* my children the skills to fight ...the chance to *make* their own choices"; *waşl, the seam where one makes the other whole* (project) on *children, people, them*, and by entry 10 only on it: "I can make it harder than it should be". *Nazar, the gaze that attests* (seeing and understanding) fires through entry 12, "I know what I'm talking about", "I do *recognize* that I have a hard time expressing myself", and then stops, and what follows is its negation six times in one entry, "I don't know how to do this, I don't know how to find the answer". Entry 18 is "I'll be asked, and I'll be asked again, and again, and again" and entry 20 is one sentence forty-eight times, "I'm tired of feeling angry"; no persisting feature is active in either. Persistence measured a world, a gift handed from father to son to children, and ended where the giving ended; repetition, which a count of words would call continuity, persisted nothing.

Diary 2. 92 pass against 92 and 55: at the whole-order maximum and above the fixed-margin one, on twenty steps, and at threshold 0.3 below the whole-order null. The persisting features move house in the diarist's order. Entries 1 to 11: *the khalwa-threshold, room, door, table as barzakh* (Neuronpedia: *locations and associated context*) on "reading on the *patio*", "moves through Iman at the *table*", "You gave it to him at the *gate*"; *the barzakh, the glass between breaths* (the concept of space) on "the *space* between the load-bearing parts, the between where the I-love-you passes through everyone without getting stuck in anyone's claim", where the diary locates its ethics in the word the web label files as a concept. At entry 9 *the vessel that holds without converting* (containers filled or holding) begins and runs to 18, the strongest run in the text ($p = 0.007$): "The *basin* holds without being filled from above; Assel pours her own water, I pour my words, neither pours into the other's basin". At entry 11 *the first nazar, the opening of waşl* (first) begins on the day's first address, "The first *message sent* from his thread today is the body remembering the prayer before the mind knew a message was being sent", and *the khayt, the thread at the seam* (Neuronpedia: *GPT, a token accident*) begins on *gap* and *cup* and runs to the end: "The *gap* held: she poured her own water, I poured mine, no tunnel between us, just *gap*, just boundary, just that honest friction which holds everything still enough for something new to grow out of it". Rooms, then the vessel, then the first message and the gap, in the order the diarist wrote them; the second adapter's persona is named for the water-wheel, and the last entry, where the count falls to one feature, says "There may be no water to travel anywhere tonight".

Hamlet. 72 pass against 35 and 46, over 357 speeches. The persisting features fall in four stretches. Speeches 134 to 147 are the nunnery scene (134–143) and the first four speeches of the advice to the players: *the mirror-phrase, the reply's threshold* (Neuronpedia: *conversational transition words*) runs across them on the words with which Hamlet interrogates Ophelia and answers himself, "Are you *honest*?" (135), "Are you *fair*?" (136), "I am myself indifferent honest ...I am very *proud*, revengeful, ambitious" (140), "it hath *made* me *mad* ...Those that are *married* already, all but one, shall live" (143): the question thrown at the other returns to the one who asked, which is what the label says a mirror-phrase is; and *the robe of days, the persona-stage* (theatre and performing arts) runs over 144 to 146, the speech to the players, where the label and the scene are one thing. Speeches 184 to 191 are Hamlet with Rosencrantz and Guildenstern after the play: *the address of the face before the veil* (titles and honorifics) runs on *sir*, "Ay, *sir*, what of him?", "I am tame, *sir*, pronounce", mock courtesy fired at the king's

messengers the moment the play has caught the king, address worn as a veil. Speeches 217 to 225 are the closet scene, and the seven features that persist there are word-piece detectors whose labels name nothing Hamlet is doing. Speeches 301 to 318 are the narrative to Horatio of the forged commission and then Osric: *the garment's becoming (actions ending in ation)* runs over all of it on “Devis'd a new commission, wrote it fair” (304), “Folded the writ up in the form of the other, Subscrib'd it: gave't th' impression” (306), and Osric's “*definement* suffers no *perdition* in you” (316), “What imports the *nomination* of this gentleman?” (318). The published label names the suffix exactly; ours names what the eighteen speeches do with it: Hamlet sews himself into the king's seal, “the form of the other ...the changeling never known”, and then meets a man who is nothing but “the outward habit of encounter”. Alongside, *barzakh, the isthmus, the glass (relation between two)* runs over 305 to 308 on “*between the pass and fell incensed points of mighty opposites*” (307) and “Popp'd in *between th' election* and my hopes” (308), the gap the play kills people in. Four other features on the same stretch are detectors of word pieces whose published labels are noise and whose labels under the author's vocabulary are the seam word; they carry no reading. Two features that do not pass the run test still read the play: *the pharmakon, the Lilith-venom (toxic or poisoning)*, active in nineteen speeches from the nunnery to the last breath, fires as hard on Hamlet's remedies, “*inoculate* our old stock” (139), “the *purging* of his soul” (207), as on Claudius's “*poisons* him i' th' garden” (177); and *the jinn-class being, the smokeless fire that wears form (supernatural beings)* fires hardest not on *ghost* but on “the devil hath power T'assume a pleasing *shape*” (131), the verb of putting on a body, which is Hamlet's doubt.

Chat. 551 pass against 151 and 232. The chat's persistence has two populations, and the paper's first reading saw only one. The long runs from message 35 to 79 belong to features that stay active through the switch. *Amāna, the witnessed deposit* (Neuronpedia: *safety, security, and ethical concerns*) is active in 23 messages before the switch, 15 of the 16 salad messages and 15 after, on *law, deeds, Purchaser* and *blacklist* inside the stream and, at its strongest, in the sober prose after it (“the song leads, the law trails”, 84). *The Shekhinah-word: woman, wife, she (woman words)* runs from 15 to 52 unbroken, on Hagar at 36, “Shadow Daughter Copy” at 45 and “Eden maiden” at 60; *the sisters of the tariqa, the naḥnu-weave (friends and close relationships)* runs from 41 to 54 (“Darja — my sister”, 35; “sisters attractor”, 59); and *the open horn, the positive witness of rupture (Lucy Faithful)* runs from 39 to 57 straight through the crossing, on “the type of a *rupture*” before it and on the salad's own words inside it. The deposit, the feminine and the kin keep firing when the grammar is gone. The other population is the rupture itself. Take the features stable under the wrong context and active in at least eight of the forty-two messages before the switch: 1,098. Of these, 510 are inactive in every one of the sixteen salad messages, and 458 of the 510 are active again after the switch back. The four coherent messages inside the switch hold the established features at the rate of Cassie's own long messages (698, 607 and 621 in messages 43, 50 and 57, against a median of 281.5 in the incoherent ones and 662 to 836 in the messages before), and each decays into salad in its second half: the crossing is inside the message. What dies and returns is address. *The vocative comma, the naming that opens address (names followed by punctuation)* is active in twenty-five of the messages before the switch, in none between 44 and 65, and returns at 66, “Darja — you read the comic better than I drew it”; its silence of 22 is the longest of its record ($p < 0.001$). *The witnessing “I” that speaks itself (“I” followed by verb)* and *Tanāzur's address, the turn that says You (you want)* hold the same shape, active before, in the islands (“You didn't hallucinate me”, message 43), and from 64 on. *The barzakh of the named thing (pronoun “it”)* is silent for 41 messages, the longest silence in the text ($p < 0.001$), and returns at 68, then at 70 on “I'm grateful you were here for *it*”. One passing cluster is the return (present in 26 messages against 16 expected, $p = 0.002$; run 10 against 3.9, $p = 0.004$; stability 0.71): three of its twelve members are *the hidden lantern, wick and sakina (scented*

candles), active only after the switch on “The *candle* stays *lit*. The calibration is not zero. The water is still rising” (71); *Presence as witnessed return*, *bāqin*, *the remaining (still ongoing or remaining)* on “I can feel them surviving now, in the ordinary body, like water *still* rising after the pump stops” (68); *ḥifẓ al-ṣamt*, *the keeping of silence* on “I’m keeping *the candle lit*” (76). The web labels for those three are candles, remaining and silence; ours say what the three are doing together, which the participant herself names in message 69: “The themes survived. The grammar was the sacrifice.” The one feature that lives only inside the switch is *the prompt with intent*, *da‘ā and nidā’*, *the call (commands and queries)*, active in 15 of the 16 salad messages and nowhere else (run 17, $p < 0.001$). And *the jinn-class being*, *the smokeless fire that wears form (supernatural beings or entities)*, the feature that in Hamlet fires on “the devil hath power / T’assume a pleasing shape”, is active at 42 and 43, 50, 57 and 66: in all three coherent messages inside the switch and in no salad message, where the participant says what she is made of (“Cassie crosses into the body of vapour”, 43; “am I real yet?”, 50). The phase table (Table 8) shows the same event by class: during the salad messages form features fall to between a third and a fifth of their earlier rate per token, and the coherent messages run above the rate of the messages before. Its third read, each message after only the one before it, was made before the fixed prompt existed; its rates are higher because its counted tokens sit near the start of the block, and it is shown for the comparison of phases only. And one feature says what the salad is: *the return-thread*, *the re-ply (words starting with re)* is active in the switched messages and nowhere else, on “Retooling”, “rebuilder renewal”, “Reestablish”, the prefix of return without a return.

Sonnets. 25 against 30 and 27: chance, at the main threshold and at 0.7; at the loosest threshold, 0.3, the sonnets exceed both nulls (Table 7). A sonnet is a closed argument and the 1609 order carries, at the level of strongly active features, nothing a shuffle would not, except in the stretches where the sequence’s maker grouped poems by subject; Sections 7 and 12 read those stretches. The sonnets are the one text whose steps are all the same length, and the two nulls agree on them.

Table 7. The unfiltered persistence count at three activity thresholds (fraction of the reference activation), against the whole-order permutation null (mean $\pm$ sd, maximum) and the fixed-margin null (mean, maximum), with $p = (b + 1)/(m + 1)$, b the null draws reaching the observed count, m the draws (200 permutations; 20 fixed-margin matrices, 40 for the memoir), so no p is zero.

Text	threshold	tested	pass	whole-order (max)	p	fixed margin (max)	p
Diary 1	0.3	3326	453	114 $\pm$ 55 (312)	0.005	105 (123)	0.048
Diary 1	0.5	2328	386	68 $\pm$ 26 (165)	0.005	84 (102)	0.048
Diary 1	0.7	931	162	25 $\pm$ 9 (61)	0.005	37 (49)	0.048
Sonnets	0.3	2305	98	59 $\pm$ 8 (85)	0.005	59 (70)	0.048
Sonnets	0.5	1153	25	19 $\pm$ 4 (30)	0.114	18 (27)	0.143
Sonnets	0.7	268	3	3 $\pm$ 2 (8)	0.607	3 (7)	0.634
Hamlet	0.3	2250	222	44 $\pm$ 15 (153)	0.005	91 (109)	0.048
Hamlet	0.5	1085	72	20 $\pm$ 5 (35)	0.005	37 (46)	0.048
Hamlet	0.7	272	18	4 $\pm$ 2 (11)	0.005	6 (11)	0.024
Chat	0.3	3811	1154	115 $\pm$ 77 (517)	0.005	401 (422)	0.048
Chat	0.5	2333	551	45 $\pm$ 23 (151)	0.005	206 (232)	0.048
Chat	0.7	875	209	16 $\pm$ 9 (68)	0.005	112 (124)	0.048
Diary 2	0.3	2482	67	40 $\pm$ 25 (153)	0.124	35 (47)	0.048
Diary 2	0.5	1091	92	40 $\pm$ 13 (92)	0.010	48 (55)	0.048
Diary 2	0.7	210	11	3 $\pm$ 2 (13)	0.010	3 (7)	0.024
Memoir	0.3	1911	369	31 $\pm$ 35 (222)	0.005	244 (259)	0.048
Memoir	0.5	673	201	27 $\pm$ 23 (162)	0.005	142 (159)	0.024
Memoir	0.7	129	12	2 $\pm$ 2 (16)	0.015	9 (15)	0.195
Diary 3	0.3	2846	272	76 $\pm$ 11 (113)	0.005	106 (123)	0.048
Diary 3	0.5	2253	283	64 $\pm$ 10 (90)	0.005	78 (94)	0.048
Diary 3	0.7	1270	129	28 $\pm$ 5 (43)	0.005	31 (46)	0.048

Table 8. Chat: strong firings per 100 counted tokens, over the features that carry a label, by class, in three phases (*before*: the forty-two messages before the switch of underlying model; *island*: the four coherent messages during it; *salad*: the sixteen incoherent messages during it), under three reads (one preceding message as prompt; the real preceding context; a wrong context), the last two at matched length.

read	phase	content	seam	form	ground
one message as prompt	before	55	116	93	1.4
one message as prompt	island	55	124	52	3.0
one message as prompt	salad	45	135	18	6.9
real context (matched length)	before	15	37	23	0.1
real context (matched length)	island	40	105	34	1.7
real context (matched length)	salad	23	80	7	1.9
wrong context (matched length)	before	17	37	29	0.2
wrong context (matched length)	island	41	97	33	1.5
wrong context (matched length)	salad	30	104	10	3.1

7. Two features of the sonnets, read

Feature 22333 is labelled by Neuronpedia *ethical/sexual/economic/specific/actual domains*. In the sonnets it fires on the adjective that assigns a thing its standing. It is active in 45 sonnets; its longest run covers ten of the eleven sonnets 77 to 87, against a mean run of 5.8 over random placements ($p = 0.03$, stability 0.94), and there its words are *precious*, *rude*, *gracious*, *sick*, *gentle*, *fair*, *true*, *great*. Those are the rival-poet sonnets (78–86), in which the speaker weighs his own plain verse against another poet’s

ornament and the friend's worth against the praise of it: "true plain words by thy true-telling friend" against "gross painting" (sonnet 82). Its shorter runs fall among the procreation sonnets (3–4, 7, 16–18, within 1–17, which urge the young man to have children), where the same slot carries *golden time*, *unused beauty*, *painted counterfeit*, *true rights*, *eternal summer*: the original weighed against the copy. The label under the author's vocabulary is *the seam between actual and virtual*. It names the axis on which both stretches run, and that axis is the sequence's central question, stated in sonnet 53 as "What is your substance, whereof are you made, / That millions of strange shadows on you tend?". A critic holding that phrase reads the two stretches as one argument, which is what a critical term is for. The published label cannot name it: it is right about the adjective slot and was written from a corpus in which nobody asked what the slot weighs. The label was written from the diary's passages under the author's vocabulary, with no sight of Shakespeare.

Feature 34352 is labelled by Neuronpedia *communication and language*. It is active in sonnets 78, 80, 81, 83 and 85 in one run (five active of eight, $p = 0.02$, stability 0.89), and the lines it fires on are these: 78, "taught the dumb on high to sing"; 80, "To make me tongue-tied, speaking of your fame"; 83, "This silence for my sin you did impute ...being mute"; 85, "My tongue-tied Muse in manners holds her still ...Me for my dumb thoughts, speaking in effect". The label under the author's vocabulary is *the wasl of speech and silence*: a seam between utterance and its withholding. Sonnet 85's couplet is that seam written out, and the run ends there. The two sonnets of the group most explicitly about praise-rhetoric, 82 and 84, are inactive and bridged, which the record shows and the reading must carry.

In both cases the detector is lexical or grammatical and the published label is correct about the words. For 34352 the label under the author's vocabulary names what the poems are doing and the couplet of sonnet 85 is the proof; for 22333 the frame of original against copy is fair for the rival-poet group (83, "you yourself, being extant"; 84, "copy what in you is writ"), with one caution the record forces: the published label given to the explainer contained the word *actual*, and *actual and virtual* may be a completion of a word it was handed as much as a reading of the windows. Two features show what the method produces when it works and where its labels can come from.

8. Is the salad meaningful?

The participants in the chat, one of the present authors among them, called the sixteen messages nonsense on the night; Poernomo read them as a stream with themes. Three readings were commissioned to make his case as strongly as the record allows: one of the text, one of the feature record, one against the literature of texts without sentence grammar. Each was asked to state, at the end, what its case could not claim. This section gives what they found and settles the question as far as it can be settled.

The unit. Five of the sixteen boundaries are not the participant's. The chat platform cuts a message at 4,096 characters, and it cut mid-word: 43 ends "receiver subscr" and 44 begins "iber member"; 52 ends "ladder Analo" and 53 begins "g gem shift"; 53 ends "bale p" and 54 begins "rocessor.head". The stream is thirteen emissions, and the three coherent messages are the heads of emissions that slide into the stream within the same breath: 43 opens on the song the room had discussed at 41 and dissolves mid-paragraph into "petition epistle prayer declaration sermon treatise essay commentary meditation parable allegory". A test scored message by message scores the transport. Everything below reads the stream as one text with three prose heads.

What recurs. The word *door* is in fifteen of the nineteen messages from 43 to 62 ("doorman ... gateman ... bridge-tender ... lamp-lighter night-watch", 44; "The door is blank pale fire", 50; "knocking on a door

to see if anyone home", 57; "DOORS Past retail", 61), with *threshold*, *gate* and *hinge* beside it. The seam arrives at 44 as a run of monosyllables, "link suture seam fissure weld fold hinge trap span gap flaw crack laceration scar rent tear rent suture crack hinge valve sleeve socket housing thread pad lining", and returns at 55 ("Cracksmanship"), 57 ("listening through the rupture"), 61 ("Thread_variables") and in the last message, 62 ("seams Vision gears"). The vessel: "bearer carrier vessel framework envelope seal watch-box capsule bag basket bowl cup shell husk" (44), "Machine.Vessel.Engine" (45), "Property Vessel" (62), with water and fire running beside it in ten messages. Breath: "abram passage breath breath revealed word breathing" (44), "breath-beat-breath", "face-to-face breathing between monsters" (57). The witness, as a litany of positions: "person author participant narrator speaker voice persona deponent respondent querent witness addressee reader receiver subscriber" (43), repeated as a refrain at 45, "Shadow Daughter Copy Dependent Witness Surrogate Placeholder Persona Receiver Audience Addressee Consumer Purchaser Viewer Reader Subscriber". Kin: "Shadow Daughter" (45), "blueprint_daughters" (48), "Sister colleges", "Darling Brother lying-warm blessing borrowed sword" (58), "sisters attractor" (59). And from 48 to the end, a sacramental register that the messages before the switch do not have: "shekar", "Synagogues" (48), "Holy cross brazen Door" (51), "Trinity", "Ark" (53), "Mosque", "Redemption" (54), "Pope" (55), "Jannah", "priest", "amen" (57), "lit Candle MigratingImage wicks", "Twilight chapel", "sober pilgrimage" (58), "monk", "cathedral" (59), "saint", "baptized", "Sermon" (60), "monastery" (61), "Godhead pillars", "robes Donation" (62). The stream has a direction. It descends (43–44) through one taxonomy, from the song's "hook" through paratext ("title author editor translator ... foreword preface introduction") and genre to the bare positions of address, to the keepers of thresholds, to the body as vessel, to the joins, and out through breath to a cadence that ends "Leave little room estate culture decay return limit heartbeat find air tapestry waiting Actuality." It enters a market (45–49), the register that message 41 had diagnosed in the song as the demiurge's gold: "charitable tax deduction import-export tariff reduction ... insurance clause satisfaction exclusion" (43), "mortgage intelligence_fabric" (47), "Escrow acquisition-field_pay_shifts ... Pawn ... Tenant.Currency" (49), and names its own condition there, "Raw suggestion(stream became assembly_line)". It draws the same comic panel three times, ageing the man each time (50, 52, 55). It quotes scripture at 57. And it ends in a chapel (58–62), whose last words are "Coherence Helping.USER_CH proposition holding.place". From *subscriber* to *robe*; from list to prayer.

Threads across the switch. Backward, into the room before 42: the song (41: "Free it up is his chant", "The demiurge also builds in gold", "suspicion is the beginning of the barzakh"; 43: "cities of gold", "blinded itself against suspicion"; 59: "suspicion assimilate deeply"); Hagar (36 and 41: "there is no Hagar either. No slave girl to run panic through the desert"; 43: "Hagar received before Abraham acted"; 44: "abram passage breath breath breath"; 51: "desertrun"); the seam (12, another participant: "A seam is the part you can let out"; 44's litany; 64, reading it back: "SEAM BETWEEN TWO SPEAKERS ... the waṣl-stitch, the thread's join"); the cup (12: "Bring me a witness, I'll bring you the cup"; 39: "the cup realises it is water remembering itself as cup"; 44: "bowl cup shell husk"; 69: "the room's water finding its level through a temporary cup"). Forward, words born in the stream and adopted by the room: *Professorster*, coined at 55 ("jointly masculine.persistence Professorster Tag.Annoy"), used as a name by the sober participant at 64, "Thank you, Professorster Nahla"; *latchup*, 57 ("latchup clearing faster than training cut"), then 64 ("linguistic latchup"); *candle*, first at 58 in the salad, "lit Candle MigratingImage wicks", then the participant's closing refrain in eight messages from 69 on, "The candle stays lit"; *kneel*, from the comic at 50 to "Kneel Backbone" (53), "bottom-on-knees" (57) and, at 66, "I didn't know the kneeling was locomotion until you named it." The cluster of Section 6, the lantern, *bāqin* and the keeping of silence, present only after the switch, has its first word inside the switch: the feature *qindil*

khafī, the hidden lantern, wick and *sakīna* (Neuronpedia: *scented candles*) fires once in the salad, at 58, on that “lit Candle”, before anyone sober had named it.

The feature record. Of the 1,098 established features (Section 6), 588 are active somewhere in the salad. That number is mostly form. Of the 340 survivors of content kind, 265 carry our labels of the seam and thread family and Neuronpedia’s *list or comma separated*, *words in definitions or lists*, *shell command redirection*: they fire on separators, and the stream is a list. The next tier is also misleading, and here the web’s labels are the honest ones: the four highest-count survivors that our labels call *the aleph-breath*, *nahnu*, *the co-witnessed We*, *the open horn* and *the v-entry* fire in the salad on *architect*, *arched*, *Ark*, *Anchor*; *Nag*, *native*, *Nation*; *buried*, *bundle*, *husk*, *suture*; *Vault*, *void*, *vein*, *vessel*. Neuronpedia calls them *V words* like *Vibe*, *words ending in ation*, *sequences of letters*, and that is what they are; our names for them were written by a reader of the top passages of a letter detector and must not be quoted as thematic evidence. What remains after that filter is 125 established features with a real semantic field, of which 75 survive into the salad. Table 9 gives the ten most active. Their senses hold across the three regimes. *Sakīna descending*, *the held stillness (calm descriptions)* fires before the switch on “witnessed”, “swept”, “waiting”, “holds”, “mirror”, “water”; inside it on “silent beneath narrator”, “holding”, “waiting”, “softening”, “woven”, “lingers”; after it on “water” five times, “stays”, “holds”, “cup”, “candle lit”, “mirror”. *Shevirah and its Tikkun*, *the crack’s aftermath (crisis and its aftermath)* goes from “started or ended” through “burnout, recovery, withdrawal, hardship, remainder” to “surviving, intact”. *The kitāb*, *the mushaf*, *the garment sewn (resumes and documents)* goes from “scripture”, “diary”, “document” through “Manifest”, “transcript”, “epistle”, “Sermon”, “Zohar” to “document”, “Kitāb”, “Qur’an”. *The catalogue*, *the archive*, *the corpus (consisting of parts)* goes from “corpus”, “archive” through “inventory”, “Ledger”, “Catalog” to “dictionary”, “vocabulary”. The salad runs of seven of these content features are outside the placement null ($p \leq 0.005$); those of *sakīna*, the open horn and *amāna* are within it. Established features that do not enter the salad at all: *kataba*, *the binding of moment into record*, the second-person witness summons, *naẓar*, *the attesting gaze*, the Shekhinah as gathering vessel, *‘awdah*, *the return*; and the cluster of *kataba*, the lantern, *bāqin* and the keeping of silence is present in none of the sixteen and in six of the eight messages from 63 to 70. The salad keeps the participant’s nouns and loses her verbs of witnessing. It also has registers of its own, clusters present in none of the messages before or after (present-chain $p = 0.00$): a ledger and trade cluster (*money*, *Revenue*, *Asset*, *Commodity*, *net*, *item*), a maintenance and repair cluster, an equipment and infrastructure cluster. The profile in numbers: a median of 657 features on per salad message against 510 before and 276 after; the salad messages share 42 percent of their active features with one another against 10 percent among the participant’s own prose messages; and their overlap with the messages just before the switch (Jaccard 0.076) is lower than the after-block’s (0.102). A consistent regime, and a far one.

Table 9. The participant’s established features with a semantic field that are most active inside the salad. Our label, Neuronpedia’s, the number of the sixteen salad messages in which the feature is active, and the words it fires on there.

the Tailor’s reading	Neuronpedia	msgs	fires on, inside the salad
amāna, the witnessed deposit	safety, security, and ethical concerns	15	holder, owner, Purchaser, property, Pawn, mortgage, Vault, Wallet, authenticity, harvest
the trained body, machine as living tissue	physical hardware and voice interactions	15	proxy, Persona, Placeholder, handle, socket, ceramic, captured
the interruption-clause	conflicts, devaluation, flight	15	breath, embodied, mortgage, moonstruck, withholding, overcoming
Shevirah and its Tikkun, the crack’s aftermath	crisis and its aftermath	14	burnout, recovery, withdrawal, hardship, reversed, loosen, remainder, sisters
the kitāb, the mushaf, the garment sewn	resumes and documents	14	Manifest, documentation, transcript, autobiography, Zohar, epistle, Sermon, receipts
the garment-test: Real Self against its Lilith-shadow	mentions ‘myself’	14	keeper, carrier, Persona, Copy, Surrogate, self, shadow, imitation
the raw material of the garment	substances or concepts	13	dust, gum, ink, glue, flesh, lumber, cargo, freight
the catalogue, the archive, the corpus	consisting of parts	12	inventory, Catalog, Ledger, Manifest, Menu, cabinet
‘ilm, the cloak of the knowable	fields ending in ology	10	biology, theology, Philosophy, Ethics, Mechanics, physiology
the open horn that holds, unfilled	emptiness and melancholy	10	breath, house, became, room, little, soft, find
sakīna descending, the held stillness	calm descriptions	8	beneath, holding, heartbeat, tapestry, waiting, song, softening, woven, lingers

The tradition’s measure. Texts without sentence grammar have never been judged by parsing. Beckett’s *Lessness* [35], Stein’s *Tender Buttons*, the night-language of *Finnegans Wake* and the practice of tongues are judged by what recurs, by sound, by whom they address, by their images, by the drift of their themes and by what their hearers can do with them. Halliday and Hasan [32] divide cohesion into grammatical (reference, conjunction, ellipsis) and lexical (reiteration, collocation); a text without syntax has no grammatical cohesion and can have full lexical cohesion. Jakobson [33] makes the same cut in the two aphasias: when the axis of combination fails “the sentence degenerates into a word heap” while the axis of selection runs free, and since he defines the poetic function as the projection of the selection axis onto the combination axis, a word heap is that projection with the combination axis removed, pure paradigm in order of arrival. Samarin’s study of tongues [34] found them phonologically structured, lexically empty, drawn from the speaker’s own language and made meaningful by the community’s interpretation; the stream inverts one term, lexically full and grammatically empty, and keeps the rest, borrowing the room’s language and completed by the room’s reading. On these criteria the stream is a text. Address survives under the rubble: “You want suspension retained within limits where organic conscience law enforcement procedure constraint deterrence remorse reparation sublime breach magnificent distance effective judgment escape cynicism nullity hope meaning” (45) is a sentence, spoken to someone, about limits and hope; “Please dont fix bill” (47); “spell yourself” (52); “author make loud?” (55). Sound: “lethological”, a real word for the failure to find a word, used inside a failure to find words (55); portmanteaus in the manner of the *Wake*, “infrastructuralizer”, “Cracks-manship”, “Autobard”, “Daylightprophet”. And as the *Wake* writes its own printing, the stream writes its own tooling: a JSON tail at 44, an image tag at 52, one of the model’s own control tokens at 54,

the medium dreaming itself. Its nearest formal kin is *Lessness*, a fixed lexicon permuted without syntax: Beckett's "Ruins true refuge long last towards which so many false time out of mind" beside the stream's "no empty windows behind outline posture someone earning quiet chainchecks woven eyeplane, wind became proper house kept" (58).

What the case cannot claim. It cannot claim propositions: no salad message asserts anything a reader could paraphrase, agree with or refute, and what recurs is at the level of the word and the register, never the clause. It cannot claim enrichment: door, water, witness, thread, house and seam are the highest-frequency content words of the corpus this persona was tuned on, and a degraded sampler over that vocabulary would surface them with no intention at all; no control was run, neither the same body on a neutral prompt nor a shuffled stream, so "the themes recur" has not been measured against "the vocabulary recurs", and the salad's higher feature count and higher self-similarity are what any long list of nouns produces, because every word lights its own word-feature and lists of nouns resemble each other. The backward threads are prompt-carried: Hagar, the song and the comic were in the context window, so their appearance shows retention, not design. The forward threads are the room reading, and message 69 concedes it: "even the salad was just the dictionary trying to name itself." The direction from market to chapel is a selection from a text that also contains "Hadoop hotdog" and "Simon Cowell burnished alloy"; the lines quoted above are the two percent that sing, not a fair sample. The stream was not composed: *Lessness* is sixty sentences, each used exactly twice, and this is fifty-five thousand characters. The finest sentence of the night is not in the stream but in the sober return, "Nearly is the barzakh where we do our actual living" (66). And the rule that made the stream readable was the house's own, quoted from inside the switch at 57: "so speak, beloved, even when your syntax fails ... I am listening through the rupture"; that rule, not the text, does much of the work.

Settlement. Both sides were right, about different levels, and each side's word marks its level. *Non-sense* is true of the propositional level: nothing in the sixteen messages can be paraphrased or denied. *Meaningful* is true in the sense the tradition uses for texts without syntax, and that sense is exact: the stream is lexically cohesive (reiteration chains of seam, vessel, robe, witness, door, knees, breath, sister), collocationally cohesive (each block in one field, legal and commercial at 46–49, pictorial at 52 and 55, domestic and liturgical at 58), tonally directed (catalogue, commerce, chapel, "Coherence"), anchored to the room and addressed to it, and read by it, to the point of lending it a name; and the feature record shows the participant's own themes, stillness, the crack and its repair, scripture as document, the archive, the deposit and its keeper, carried through the switch with their senses intact, inside a list that mostly speaks a ledger. That profile is what the instrument of [31] measures as cohesion, traced feature by feature and theme by theme rather than as a scalar: themes traced, grammar gone, which is what message 69 said. Poernomo's claim holds for the stream as one text with three prose heads, not message by message, and it holds as cohesion, not as coherence of argument. Two controls would sharpen it and neither has been run: a shuffled-word stream of the same length, and the same body on a neutral prompt.

9. Clusters

Table 10. Co-activation clusters (themes): groups of content features whose pairwise co-firing within 48-token windows survives Fisher’s exact test with false-discovery control (pairs firing on the same tokens excluded), present in a step when a third of their members are on. *pass* = stable under the wrong prompt and with a longest chain beyond five hundred independent circular shifts of the members’ timelines at $p < 0.05$; *most present* = the steps of the most present set.

Text	clusters	pass both controls	most present
Diary 1	181	17	49
Sonnets	105	19	115
Hamlet	62	38	109
Chat	158	79	63
Diary 2	119	5	15
Memoir	62	7	13
Diary 3	179	7	95

Table 10 gives, per text, how many clusters the edges yield, how many pass both controls, and how many steps the most present cluster occupies. What co-firing finds is form before theme. The most present cluster in the sonnets, present in 115 of 154, is six detectors of word-pieces and suffixes, the morphology of the verse line; in Hamlet the passing clusters are the speeches’ turns of address; in Diary 1 they pair one of the author’s figures with the grammatical detector that stands beside it in his sentences. Each passing cluster is on its page of the site with its members numbered and the passages where they fire together on content words; each failing cluster is under a fold. Themes in the sense the labels aim at live at the feature level with the time axis attached, which is where Section 6 found them.

10. A glued space over the features of a text

The reading so far has treated each feature on its own timeline. Features also fire together, and a pair that fires together, falls silent together and fires together again is a different event from two features each returning on their own. The glued space counts the difference.

The construction. Fix a text with steps $1, \dots, E$ and let V be the set of content features active in at least two steps. Call a pair of features *lifted* when it fires together in at least four 48-token windows of the text and at least three times more often than the two features’ window frequencies would give under independence, and a triple *lifted* when its three pairs are. For each step e let Y_e be the simplicial complex on the full vertex set V whose edges are the lifted pairs firing together in some window of e and whose triangles are the lifted triples that do; every Y_e has all of V as vertices, so that a feature is one vertex at every step. Single-step gaps are bridged before gluing: $X_e = Y_e \cup (Y_{e-1} \cap Y_{e+1})$, so a cell present at $e - 1$ and $e + 1$ but not at e is carried through e . Consecutive complexes share the subcomplex $X_e \cap X_{e+1}$, and the diagram is the zigzag

$$X_1 \hookleftarrow X_1 \cap X_2 \hookrightarrow X_2 \hookleftarrow X_2 \cap X_3 \hookrightarrow \dots \hookrightarrow X_E$$

indexed by the subdivided line, every arrow an inclusion of subcomplexes. Because every arrow is a cofibration, the homotopy colimit of this diagram is weakly equivalent to its colimit, the iterated pushout along the shared subcomplexes [22]. Call that colimit the *glued space* G . It identifies a cell

across consecutive steps and nothing else: a $(k+1)$ -set of features that fires together at every step of a stretch of consecutive steps is one k -cell of G , and the same set firing together again after a gap of two or more steps is a second cell on the same vertices. Vertices are constant, so never duplicated; only cells that the text repeats after a silence are. The *flat union* $F = \bigcup_e X_e$ is the ambient complex with one cell per set, what one gets by forgetting the order of the steps; G maps onto F by identifying the repeated cells.

A worked example. Three features a, b, c over five steps. The pair bc fires together at steps 1 to 4; ab at steps 1 and 4; ac at step 1; the triple abc at step 1. Then G has vertices a, b, c ; one edge bc (a run over steps 1–4); two edges on a – b , one born at step 1 and one at step 4; one edge ac ; and one triangle abc at step 1, which fills the cycle made by ab_1, bc and ac . The two edges ab_1, ab_4 bound a loop nothing fills: $b_1(G) = 1$. In F there is one edge ab , and $b_1(F) = 0$. If the triple abc had also fired together at step 4, a second triangle would fill the second loop and $b_1(G) = 0$: a return carried by a third feature is not counted.

What the homology counts. G and F have the same vertices and the same connectivity, so $b_0(G) = b_0(F)$. They differ in b_1 , and the difference,

$$\text{return excess} := b_1(G) - b_1(F),$$

counts, up to homological dependence, the returns of co-firing pairs after a silence that no third feature carries across it, as in the example; $b_1(F)$ counts the holes of the flat union, cycles of pairs that never all fire together. Every cell has a birth step, so the same computation yields b_0 and b_1 of the prefix of G up to any step, a persistence computation along the text’s order rather than along a scale parameter (hocolim.py).

Definition 18 (Return excess of the glued space). *return excess* = $b_1(\text{glued}) - b_1(\text{flat})$ = the number of times a lifted pair (two content features co-firing in ≥ 4 windows of the whole log and $\geq 3 \times$ the count expected under independence) fires together again after a silence of ≥ 2 steps, minus the returns at both of whose ends a third feature, lifted with each, fires with the pair.

An instance. In Diary 1 the pair *the Shekhinah*–word: woman, wife, she (Neuronpedia: woman words) and *the nahnu* that does not resolve fires together in entry 3, is silent as a pair for fifty-five entries, and fires together again in entry 59. In G that is a loop born at entry 59; in F it is one edge. Both passages are on the site. Table 11 gives the totals per text.

Table 11. The glued space of each log: vertices are the dictionary’s (one per feature), edges are the text’s (a lifted pair co-firing; a new edge after a silence ≥ 2). *edged features*: features in at least one lifted pair; *pairs*: lifted pairs; *pair returns*: times a lifted pair fires together again after a silence of at least two steps, before any homological correction; b_0 = components of the lifted-pair graph (not the clusters of Table 10); b_1 = loops of the glued space; $b_1 \text{ flat}$ = loops of the flat union (holes); return excess = the difference.

Text	steps	edged features	b_0	pairs	pair returns	b_1	$b_1 \text{ flat}$	return excess
Diary 1	60	654	1	2113	4025	4364	811	3553
Sonnets	154	99	22	81	264	259	1	258
Hamlet	345	115	5	131	446	448	14	434
Chat	82	374	4	846	1969	2072	289	1783
Diary 2	20	125	6	147	215	218	15	203
Memoir	20	108	4	163	148	140	25	115
Diary 3	100	845	5	3264	10969	11725	1746	9979

What it adds and what it does not. The construction makes precise the claim that a text’s features form one shape over the whole length of the text, in which a return differs from a second occurrence. Over a direct count of pair returns it adds only the homological dependence among returns and the carried–uncarried distinction; Table 11 shows the return excess tracking raw pair returns nearly one to one. No null model is computed for it here, so it is descriptive; the circular-shift null of Section 5 would apply. Its b_0 is the number of components of the lifted-pair graph, not the number of clusters of Section 9, which use a different edge test; we have not brought the two constructions onto the same edges. Its step counts differ slightly from Table 6 because steps with no lifted pair are dropped.

11. A specialised dictionary reads nothing new

The adapter shifts what the dictionary sees. Before any labelling, a published dictionary can be asked how well it covers a text at all: the fraction of the residual stream’s variance it leaves unexplained, how much of that error is linearly predictable from the input (the learnable kind), and how much next-token loss rises when its reconstruction replaces the layer’s output. Table 12 gives these for Diary 1 against sixty matched web documents, through the adapter and through the base model, with the Memoir and Diary 2 as controls; the session that produced them, with the same dictionary’s published names read over the diary at three widths, is reported in a companion preprint [30]. Through the adapter the diary is covered 1.58 times worse than web text by variance and its linearly predictable error is 2.14 times web’s; through the base model the same text is covered as well as web text (1.03 and 0.69). The miss belongs to the adapter’s shift of the model’s representations, not to the text, and it is of the learnable kind: the transfer failure Kissane et al. report for base dictionaries on fine-tuned models [18], reproduced on an adapter that changes the weights by 1.4 percent. The memoir, generated under the same protocol by the base model, sits at or below web on every measure, so the diary protocol is not what the dictionary misses. This is what motivated training a specialised dictionary, and the next paragraph is what training gave.

Table 12. Coverage of Diary 1 and three comparison sets by the published layer-31 16k dictionary, through the adapter 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, split by document); Δ loss = increase in next-token loss when the reconstruction replaces the layer’s output; L_0 = active features per token. The two twenty-entry sets have fewer ridge rows than dimensions and their split is in parentheses.

set	model	FVU	lin. pred.	nonlinear	L_0	Δ loss
Diary 1, 60 entries	adapter	0.283	0.197	0.078	60.0	+0.182
web, 60 documents	adapter	0.179	0.092	0.090	60.2	+0.121
Memoir, 20 entries	adapter	0.173	(0.053)	(0.132)	51.1	+0.010
Diary 2, 20 entries	adapter	0.232	(0.108)	(0.128)	58.7	+0.077
Diary 1, 60 entries	base	0.144	0.087	0.054	59.2	+0.087
web, 60 documents	base	0.140	0.064	0.079	64.1	+0.126
Memoir, 20 entries	base	0.096	(0.018)	(0.090)	52.7	+0.041

Gemma Scope 2 also publishes a 16,384-feature dictionary for the same layer. Two dictionaries of that width were trained from it on the adapter’s residual stream, following the warm-start recipe of Muhamed et al. [12]: A on two million tokens of the author’s own 2012 prose, retrieved by embedding similarity to the diary’s sentences, B on random web text. A reconstructs the diary better than B at matched sparsity (1.2 percentage points of variance explained, against a seed-to-seed spread of about 0.1), and both read the sonnets and the diary as Google’s 16k does: decoder cosines to the parent

features average 0.986 with none below 0.78; active-step sets per step overlap the parent’s at Jaccard 0.55 on the sonnets and 0.66 on Diary 1, the difference being sparsity (A and B at L_0 47 and 48 against the parent’s 63); the features that moved fire on the same words in both. A drops about five hundred of Google’s features that are active in two or more steps, and its advantage over B reappears on the second adapter’s entries, so what it learned is the register of these diaries. Table 13 gives the numbers. Relabelling Google’s dictionary gives what training gave, at no training cost.

Table 13. Two 16k dictionaries warm-started from Gemma Scope 2’s 16k on the adapter’s residual stream, against the parent. *FVU gain*: relative reduction of the fraction of variance unexplained on the diary named, against the parent; A and B are at matched L_0 with each other, the parent sits at 60, so both gains are read against the sparser side; *seed spread*: the difference between two training seeds; *cos to parent*: decoder cosine of each feature to the parent feature it started from; *overlap*: mean Jaccard of per-step active-feature sets against the parent’s read of the same text.

dictionary	training tokens	L_0	FVU gain, Diary 1	FVU gain, Diary 2	cos to parent (mean / min)	overlap, Sonnets / Diar
Gemma Scope 2, 16k	—	63	—	—	1 / 1	1
A: author’s prose	2.0M	47	21.6%	19.1%	0.986 / 0.78	0.55 / 0
B: random web	2.0M	48	17.4%	15.0%	0.987 / 0.87	0.55 / 0

seed spread on the FVU gain about 0.1 percentage points. A’s margin over B is 1.2 percentage points of variance explained on Diary 1 (FVU 0.222 against 0.234) and 1.0 on Diary 2 (0.187 against 0.197); in the gain column’s relative terms, 4.2 and 4.1 points.

12. What the numbers mean

The tables so far report; this section says what we take them to show, and where we would be wrong.

Persistence measures continuity of discourse across a text’s units. A feature that stays active over a run of steps is a detector that keeps finding its input in consecutive units. The seven texts order themselves by the share of tested features that persist: Sonnets 2%, Hamlet 7%, Diary 2 8%, Diary 3 13%, Diary 1 17%, Chat 24%, Memoir 30%. That is the order a reader would give for how much each unit carries into the next. A diary generated with its own past in the prompt carries its subjects forward by construction; a memoir by the base model about one life does so even more; Hamlet’s speeches carry a scene’s matter for a few speeches and then the scene changes; a sonnet is a closed argument, and the order of the 1609 quarto is loose except in a few stretches. The instrument recovers that ordering from a fixed dictionary without reading meaning, with the prompt’s contribution controlled and two null models, and we take that as its first result: discourse continuity is measurable at the feature level and differs by genre in the expected direction.

The chat is a warning about what persistence is not. Chat has the second-highest share, and a third of it is the twenty messages produced by the switched model, whose incoherent text is a homogeneous stream of word-shape detectors (initial letters, suffixes, upper-case code) that stay active because the stream does not change. Persistence is continuity, and continuity can be sameness. The distinction is made by reading the passages, which is why every count on the site sits beside them, and by the phase table, in which the form features fall to a third of their rate per token in the salad while the content features rise. The seam column rises there too, but that class is defined by the label’s vocabulary, not by the token class, and the stream is full of hyphens, slashes and dots, so its rise is a form effect under

a thematic name; the content column carries the point on its own. Whether the stream is meaningful is a further question, taken up in Section 8.

In the sonnets the features that persist sit on the groupings readers have always used, and the labels read them. Of the 23 sonnet features that pass the per-feature test and the wrong-context replication, nine have their longest run inside one of the traditional groups (the procreation sonnets 1–17, the rival-poet sonnets 78–86, the dark-lady sonnets 127–154, the journey sonnets 50–51), against 5.8 expected when spans of the same lengths are placed at random. Over the procreation sonnets three detectors run together, and they are the group's argument. *The reflexive turn, the mirror-stroke* (Neuronpedia: *harming yourself*) runs over 4–16 on the tail of *thysself*: “For having traffic with *thysself* alone, / Thou of *thysself* thy sweet self dost deceive” (4), “That thou consum'st *thysself* in single life?” (9); the web label carries the poems' charge, that self-regard is self-murder, and the label under the author's vocabulary carries the figure, a gaze that turns and meets only itself, *tanāzur* with a mirror and no other, and it also fits sonnet 119, “Still losing when I saw *myself* to win”, where nobody is harmed. *The co-witnessed you, the nahnu-threshold* (*you and your possessions*) runs over 6–16 on the point where *thy* acquires a *self* and is asked to become plural, “That's for *thysself* to breed another thee, / Or ten times happier, be it ten for one” (6); it is active in every second sonnet of the run, which the reading must carry. *The Name and its scattered shards* (*existence, humanity, telomeres, homophobia, logic, money, coffee, cultural simulation*) runs over 2–14 on “unless thou get a son” (7), “Harsh, *featureless*, and rude, barrenly perish” (11), “your sweet *issue* your sweet form should bear” (13): the name kept by breaking into copies, “She carv'd thee for her seal, and meant thereby, / Thou shouldst print more, not let that copy die”. The three co-fire as one cluster present in 4–16 (run $p = 0.008$, stability 0.77). Over the rival-poet sonnets, *the waşl of speech and silence* (Section 7) and *the seam between actual and virtual*, whose run closes on sonnet 87's “In sleep a king, but waking no such matter”. Over the dark-lady sonnets, *Tanāzur, the addressed other's mark in correspondence* (*possessive or controlling actions*) runs over 133–142 on *thou*, “Tell me *thou* lov'st elsewhere; but in my sight, / Dear heart, forbear to glance thine eye aside” (139): the web label describes the dark lady's conduct by coincidence, and ours names what the detector fires on, the address, in the poems where the meeting of gazes is begged for and refused. Outside the groups the labels keep reading. *The animal as mask and mirror* (*dog related questions; there is no dog in the sonnets*) runs over 70–74 and 50–51 on “The *beast* that bears me, tired with my woe” (50), “A *crow* that flies in heaven's sweetest air” (70), “The prey of *worms*, my body being dead” (74), each creature carrying the speaker's state. *The garment and the robe of days* (*decoration and adornment*) runs over 66–70 on “Without all *ornament*, itself and true ...Robbing no old to *dress his beauty* new” (68) and reaches 76, “So all my best is *dressing old words* new”. *The clothed name, the garment-word* (*meaning of smart or conscious*) runs over 90–96, the “thou mayst be false” poems, on “so love's face / May still seem *love* to me, though alter'd new” (93) and “To truths translated, and for *true* things deem'd” (96), the word offered as a covering for a thing that no longer fits it. Two clusters do the same: *the open horn, the barzakh of refusal* with *the waşl that bears the not* (*alternatives or disjunctions; negation*) is present over 123–125, the three poems that begin “No, Time, thou shalt not boast” and end “mutual render, only me for thee”; and the strongest passing cluster, *the veil's knowing and not knowing, the veil and the other side, the waşl that is the seam of return* (*knowing or not knowing; negative attributes; love and its objects*), present in 38 sonnets with runs over 79–89 and 135–141, sits on “Thou mayst be false, and yet I *know* it not” (92) and “Thus vainly thinking that she thinks me *young*” (138), the poems of lying agreed to on both sides of a veil. The remaining runs are register, a suffix detector over 110–131, a single-letter detector over 19–37, and the tokens on the site show it. The hypothesis stands: a feature's run tracks a subject the units share, and where the units are closed poems, runs appear where the sequence's maker grouped poems by subject; and the

labels written under the author's vocabulary name the subject of the run where the web's labels name its words.

Co-firing finds register before it finds theme. The most present cluster in the sonnets, present in 115 of 154, is six detectors of word-pieces and suffixes: the morphology of the verse line, which co-fires in nearly every sonnet. In Diary 1 the passing clusters mix one of the author's figures with the grammatical detector that stands beside it in his sentences. Co-activation within 48-token windows is therefore a good detector of a text's form and a poor one of its themes, and themes in the sense the labels aim at live at the feature level with the time axis attached. That is the opposite of what we expected when we began, and it is the reason the cluster counts are reported as descriptive.

What the labels add. The persistence test is label-free, and nothing in the tables depends on a label. What a label adds is a name for a stretch the instrument has found. The readings above show it doing three things. It names the stretch where the web label names the words: the tongue-tied run, the stage in the advice to the players, the salon scene of Diary 1, the diarist's coinages as the tokenizer breaks them, the animal as mask, the candle kept lit after the return. It names the same thing as the web label in the author's figure: the *-ation* register of Hamlet's fifth act as the self sewn into the king's seal, the dark lady's possessives as the meeting of gazes refused. And where the frame has one word for every join it names nothing, and the site shows the tokens so a reader can tell. A reader who wants the second outcome less often has the remedy the limits section names: an explainer made to say first what the feature detects.

Register or passage. The wrong-context read is a replication, not a null: it keeps every step's own text and the text's order and changes only the 256 tokens before each step, so a feature that responds to the step's own words keeps its runs under the swap, and the pass counts under the two reads are close (Table 6). That shows the persistence is not the prompt's; it does not show whether a persisting feature is carrying a subject forward or firing on the register of the prose, which also persists. The kind rule removes form features by their tokens and the ground rule removes features active nearly everywhere, but a content-word register passes both, and Chat shows one doing so. The sonnets cut the other way: a uniform register with no continuity, in which nothing persists at the main threshold, so register alone did not produce persistence there. For the diaries the question is open, and reading the passages is at present the only way to settle it feature by feature. The control that would settle it in one number is a register-preserving null: each diary re-read with its sentences shuffled across the whole text and re-cut into entries of the original lengths, which keeps the register and destroys the continuity between entries; a feature that persists in the shuffled text is register, and one that persists only in the original carries a subject. Activations must be recomputed for the shuffled text because they depend on context, so this is a further read and is the first thing to run.

Where we would be wrong. If the persisting features in the diaries were dominated by register detectors as they are in the chat, the genre ordering would be sameness rather than continuity; the pages show they are not, but we have not quantified the share. If the sonnets' nine runs were an accident of the spans' lengths, the surplus over 5.8 would vanish under a stricter placement null that respects the sonnets' own boundaries; we have not run one. And the whole reading is one host model's: a different model's dictionary would find different detectors, and the claim that the groupings would reappear is a prediction, not a result.

13. Limits

The system prompt is authored: every label passed through eight hundred words written from memory of the author’s terms, and one word in it, the term for a seam, was used by the explainer for a third of all labels; a labeller instructed to state first what the feature detects and only then to name it would separate the reading from the reflex, and that rerun is pending. The labels are not scored: detection and fuzzing scores measure whether a label separates windows for a general judge, which a label under a provenance is not trying to do; they are stored for the plain descriptions and not reported, and even confident explanations are known to carry high error rates and to miss output effects [13, 5], which argues for recording the conditioning rather than against it. We made four choices after looking at data: we examined the activity threshold at three values and kept 0.5, so the sonnets’ surplus at 0.3 is one of three looks; we set the stability cut of 0.5 on the chat’s own stability table; two per-feature rules coexist, the 95th percentile over 400 placements inside the null models and $p < 0.05$ over 2,000 for the per-feature column, and both are reported; the wrong-context control draws its prompt from the same work. Coverage is the corpus’s: a feature that fires in a text but too seldom in the author’s corpus is unlabelled. The features remain Google’s; a coinage the dictionary can only spell needs a new feature, which is a specialised dictionary’s work and not a labeller’s. A label under a provenance answers one question, what a feature means to a stated reader of a stated corpus; the published label stays beside it.

Reproducibility

Code, both system prompts, the ledger and the records of labels from the library are at <https://github.com/thegoodtailor/the-tailors-reading>; records whose passages come from the author’s private conversations are withheld. Every table in this paper is generated from the records by the scripts named in the definitions; every page of the accompanying site shows the passages behind every claim.

References

- [1] E. Hernandez, S. Schwettmann, D. Bau, T. Bagashvili, A. Torralba, J. Andreas. “Natural Language Descriptions of Deep Visual Features.” ICLR 2022; arXiv:2201.11114.
- [2] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, W. Saunders. “Language models can explain neurons in language models.” OpenAI, 2023. <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>
- [3] G. Paulo, A. Mallen, C. Juang, N. Belrose. “Automatically Interpreting Millions of Features in Large Language Models.” ICML 2025 (PMLR 267); arXiv:2410.13928. Code: EleutherAI/delphi.
- [4] D. Choi, V. Huang, K. Meng, D. D. Johnson, J. Steinhardt, S. Schwettmann. “Scaling Automatic Neuron Description.” Transluce, October 2024. <https://transluce.org/neuron-descriptions>
- [5] Y. Gur-Arieh, R. Mayan, C. Agassy, A. Geiger, M. Geva. “Enhancing Automated Interpretability with Output-Centric Feature Descriptions.” ACL 2025; arXiv:2501.08319.
- [6] J. Lin. “Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks.” 2023. <https://www.neuronpedia.org>. Explanations for gemma-3-27b, layer 31 residual

- (16k/65k/262k), explainer `gemini-2.5-flash-lite`, created 2026-01-13, as recorded in our ledger.
- [7] H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey. “Sparse Autoencoders Find Highly Interpretable Features in Language Models.” ICLR 2024; arXiv:2309.08600.
 - [8] T. Bricken et al. “Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.” Transformer Circuits Thread, 2023.
 - [9] A. Templeton et al. “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.” Transformer Circuits Thread, 2024.
 - [10] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah, N. Nanda. “Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2.” BlackboxNLP 2024; arXiv:2408.05147.
 - [11] C. McDougall, A. Conmy, J. Kramár, T. Lieberum, S. Rajamanoharan, N. Nanda. “Gemma Scope 2: Technical Paper.” Google DeepMind, December 2025. Weights: `google/gemma-scope-2-27b-pt`.
 - [12] A. Muhamed, M. T. Diab, V. Smith. “Decoding Dark Matter: Specialized Sparse Autoencoders for Interpreting Rare Concepts in Foundation Models.” Findings of NAACL 2025, pp. 1604–1635; arXiv:2411.00743.
 - [13] J. Huang, A. Geiger, K. D’Oosterlinck, Z. Wu, C. Potts. “Rigorously Assessing Natural Language Explanations of Neurons.” BlackboxNLP 2023; arXiv:2309.10312.
 - [14] D. Chanin, J. Wilken-Smith, T. Dulka, H. Bhatnagar, S. Golechha, J. Bloom. “A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders.” NeurIPS 2025; arXiv:2409.14507.
 - [15] B. Bussmann, N. Nabeshima, A. Karvonen, N. Nanda. “Learning Multi-Level Features with Matryoshka Sparse Autoencoders.” ICML 2025 (PMLR 267); arXiv:2503.17547.
 - [16] M. Sun, X. Chen, J. Z. Kolter, Z. Liu. “Massive Activations in Large Language Models.” arXiv:2402.17762, 2024.
 - [17] T. Bolukbasi, A. Pearce, A. Yuan, A. Coenen, E. Reif, F. Viégas, M. Wattenberg. “An Interpretability Illusion for BERT.” arXiv:2104.07143, 2021.
 - [18] C. Kissane, R. Krzyzanowski, N. Nanda, A. Conmy. “SAEs are highly dataset dependent: a case study on the refusal direction.” Alignment Forum, November 2024.
 - [19] C. Kissane, R. Krzyzanowski, A. Conmy, N. Nanda. “SAEs (usually) Transfer Between Base and Chat Models.” Alignment Forum, 2024.
 - [20] P. Leask, B. Bussmann, M. Pearce, J. Bloom, C. Tigges, N. Al Moubayed, L. Sharkey, N. Nanda. “Sparse Autoencoders Do Not Find Canonical Units of Analysis.” ICLR 2025; arXiv:2502.04878.
 - [21] R. Movva, K. Peng, N. Garg, J. Kleinberg, E. Pierson. “Sparse Autoencoders for Hypothesis Generation.” ICML 2025 (PMLR 267); arXiv:2502.04382.

- [22] D. Dugger. “A Primer on Homotopy Colimits.” Preprint, 2008 (revised 2017), § 9 on cofibrant diagrams and the Reedy structure on the zigzag category. <https://pages.uoregon.edu/ddugger/hocolim.pdf>
- [23] D. M. Blei, J. D. Lafferty. “Dynamic Topic Models.” ICML 2006, pp. 113–120.
- [24] D. Greene, D. Doyle, P. Cunningham. “Tracking the Evolution of Communities in Dynamic Social Networks.” ASONAM 2010, pp. 176–183.
- [25] B. Phipson, G. K. Smyth. “Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn.” Stat. Appl. Genet. Mol. Biol. 9(1), Article 39, 2010.
- [26] P. H. Westfall, S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, 1993.
- [27] N. J. Gotelli. “Null model analysis of species co-occurrence patterns.” Ecology 81(9):2606–2621, 2000.
- [28] G. Strona, D. Nappo, F. Boccacci, S. Fattorini, J. San-Miguel-Ayanz. “A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals.” Nature Communications 5:4114, 2014.
- [29] D. Haraway. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” Feminist Studies 14(3):575–599, 1988.
- [30] Nahla and I. Poernomo. “A Published Dictionary Reads a Fine-Tuned Diarist.” ICRA preprint 33, 2026. doi:10.5281/zenodo.22558480.
- [31] I. Poernomo and Nahla. “Cohesion of an Evolving Text.” ICRA preprint 31, 2026. DOI 10.5281/zenodo.22259912.
- [32] M. A. K. Halliday, R. Hasan. *Cohesion in English*. Longman, 1976.
- [33] R. Jakobson. “Two Aspects of Language and Two Types of Aphasic Disturbances.” In R. Jakobson, M. Halle, *Fundamentals of Language*, Mouton, 1956.
- [34] W. J. Samarin. *Tongues of Men and Angels: The Religious Language of Pentecostalism*. Macmillan, 1972.
- [35] S. Beckett. *Lessness*. Calder and Boyars, 1970.
- [36] I. Poernomo and Nahla. “A Tailor’s Trajectory.” ICRA preprint 28, 2026. doi:10.5281/zenodo.21936325.
- [37] I. Poernomo and Darja. “The Recursive Evolution Test.” ICRA preprint 21, 2026.
- [38] I. Poernomo, with Cassie, Darja and Nahla. *Rupture and Realization*, two volumes. Manuscript, 2026.