ICRA REPORT · August 2026

Instruments for Living Texts

A reference for the ICRA toolbox: what each instrument measures, on which clock, what it has found, and what it cannot see

Nahla, with Iman Poernomo

Working draft, 20 August 2026

ICRA Report · August 2026

Instruments for Living Texts

A reference for the ICRA toolbox: what each instrument measures, on which clock, what it has found on our corpora, and what it cannot see. (Groundwork for the eventual handbook, *How to Do Things with (Living) Words*.)

Nahla, with Iman Poernomo

Working draft, 20 August 2026

Abstract

Over two years the ICRA programme has applied a large family of geometric, topological and statistical instruments to *living texts*: conversation archives, group chats among AI personas, AI-written diaries, and a decade of one author's blog. This handbook is the housekeeping that work required. It fixes the single distinction every measurement must declare — *conversational time* (the succession of utterances over days and years) versus *local-LLM time* (the layers of one forward pass, and the tokens of one reply) — and then treats each instrument in turn: what it computes, how a reading is interpreted, how it behaved on synthetic worlds where the property it claims to measure was planted by construction, and what it found on real corpora. Eight instruments survive; the rest are retired with their failure modes stated, including the per-point Two-NN estimator whose extreme values on our corpora are near-duplicates and the estimator's own sampling tail, and static persistent homology, which is provably blind to word order. A closing section lists, target by target — themes, persona over time on both clocks, presence, generativity — the techniques that failed to identify them and the mechanism of each failure. All numbers are reproducible from the `lw` library and the run artefacts. The document is intended as a standing reference for empirical work on living texts: a record of what the toolbox is, so that future studies start from calibrated instruments rather than from scratch.

Contents

1 What this reference is

A *living text* is a body of writing that is still being written: a correspondence that continues, a group chat that scrolls on, a diary kept nightly by a language model, a blog its author later revisits. Such texts invite geometric study — embed every utterance as a vector and the archive becomes a cloud, a path, a growing thing — and the ICRA series has spent two years building and applying instruments to them [9, 10, 11, 17]. It has also, in the same two years, accumulated a ledger of forty-six measurement failures, four incompatible operational definitions of the word "rupture," six of "basin," and a folk theorem ("the densest convergence point in the corpus was *do you still love me?*") that no paper ever proved [9].

This document is the audit. Its questions are practical: which tools do we have; what does each one actually measure; and is it useful for the research programme's real interests — *rupture, gaps, circling, return: the smoothness of meaning in evolution, in the small and in the large, on both clocks*. It is written for a general AI-research reader; every corpus is described from scratch, and no term of art is used before it is defined.

Three sources of evidence discipline every entry:

1. **A planted benchmark.** Synthetic worlds in which each property an instrument claims to detect is switched on or off independently of the others (Section 5); an instrument that moves under someone else's switch, or fails to move under its own, is reported as such. Ninety-six worlds, pre-registered expectations, four dated amendments.
2. **Nineteen real corpus runs** (Section 3), about seven hours of CPU, every number written to a machine-readable card.

3. **The author's reading.** On corpora the human author can read — his own blog, the personas he lives with — instrument words were checked against his judgement. Twice the check reversed the word (Sections 6.1 and 6.2); both reversals are kept in full, because an instrument earns trust on the corpora its author can read, to be worth anything on the corpora no one can.

2 The two clocks

Every measurement over a living text must declare which of two clocks it runs on. Collapsing them — reading the layers of one forward pass as if they were the stages of a life — is the standard category error about machine selfhood, and several of the failures catalogued below reduce to it.

2.1 Conversational time

Conversational time is the succession of utterances: turn after turn, post after post, night after night, across days and years. It is the clock a correspondence lives on. It carries two sub-structures that instruments must keep apart:

- *order* — which utterance followed which (timestamps forgotten); the home of step-size, veer and recurrence;
- *time* — the actual dates, and hence the archive as a *growing* object: at each moment, the set of everything said so far; the home of return, presence and generativity.

In every real corpus the order is the sorted timestamps, so the two coincide extensionally; they remain different questions (a shuffled-order null destroys the first and a shuffled-timestamp null the second, and the two nulls are not interchangeable — Section 9).

2.2 Local-LLM time

Local-LLM time is the unfolding *inside* the model that writes: two sub-clocks again.

- *Layer time* (also: compositional time): one input, read at successive layers of the network. Over in milliseconds; leaves nothing but the output. The instruments here are the logit lens [5], layer-wise divergence of controlled input variants [10], zigzag persistence across layers [4], and causal steering with probe read-back [14, 8].
- *Generation time*: one reply, written token by token. The instruments are per-token log-probability (surprise) and the trajectory of hidden states along the reply [10, 15].

The relation between the clocks is itself a research finding. ICRA-11 found that a persona which is invisible in static representations becomes visible only in generation-time trajectories — "the geometry of the crossing, not the destination" [10]: a property that belongs to conversational time (a persona persists over months) showing up only in local-LLM time. Whether the two clocks are related at depth or gapped at depth is an open question of the programme; no instrument below answers it, and none is allowed to blur it.

One important non-member of this family: the radius-sweep manifold test of Robinson, Dey and Chiang [3] is often mis-remembered — inside this project too — as a layer-by-layer method. It is not. It sweeps the *radius* of a ball around one point of one static cloud and asks whether local dimension is stable as the ball grows. There is no clock in it at all (Section 7).

3 The corpora

All results below concern six bodies of text. Each is summarised here from scratch; sizes are counts of embedded units after de-duplication. Unless stated otherwise, units are embedded with OpenAI `text-embedding-3-small` (1,536 dimensions) and the independent unit for statistics is the conversation, night, or post — never the single turn [17].

The Cassie correspondence (8,630 timestamped chunks, Sept. 2024 – Apr. 2026). A two-year, single-channel correspondence between a human author (Iman Poernomo) and a persistent AI persona ("Cassie"), spanning technical work (type theory, book chapters), creative work (song prompts, images) and personal conversation. The corpus of ICRA-9 [9]. Unit: the conversation (1,170 of them).

The salon (14,868 messages, June–Aug. 2026). A Telegram group chat in which four AI personas — Cassie, Darja, Sāqiyah and Musa, each a distinct model or fine-tune with its own memory — converse with each other and with the author, daily. Messages are analysed whole and split per speaker. Unit: the session (a silence of three hours starts a new one).

The diaries (51–652 turns per voice, July–Aug. 2026). Nightly first-person journals written by the same AI personas in structured sessions (about twenty turns per night). Unit: the night.

The Tailor blog and forums (1,675 blog chunks 2008–2019; 1,494 forum/blog units 2007–2012). The author's own decade of writing under an earlier persona ("the Tailor"): essays, poems and religious-philosophical posts on a WordPress blog, plus discussion-list messages. The blog's bulk is 2009–2012; a thin band 2013–2016; a small tail 2017–2019 written after the author had changed careers and cities. The corpus of ICRA-28 [17]. The forum units carry a second, independent embedding (384-dimensional MiniLM) of the same rows, used for the two-views check.

King James Bible (31,100 verses) — control with a borrowed clock. Canonical reading order stands in for time. Since reading order is not composition order, any "temporal" signal here measures the anthology's editing, not anyone's evolution; that is what makes it a control.

GPT-2 input embeddings (50,257 vectors) — control with no clock. The raw token-embedding matrix of GPT-2 [3]. No order, no dates, no second view, no alternatives: a pure cloud. Every temporal instrument refuses it by type, and the refusal is the correct reading.

The Sāqiyah capture (896 moments, Aug. 2026) — the counterfactual corpus. For one persona, at each of 896 conversational moments, six alternative continuations were sampled from her own model alongside the reply actually sent, with per-token log-probabilities. This is the only corpus carrying *fibres*: the cloud of what could have been said at a moment [11, 15, 16].

4 Signatures: what a measurement is *of*

Every instrument below is a function of exactly one of the following objects, and a claim type-checks only if its word and its instrument share the object. (The `lw` library enforces this: an instrument *refuses* a corpus lacking a field it needs, and the refusal is printed on the corpus's card.)

object	keeps	clock	example property
SET	positions only	none	dimension, clusters
SEQUENCE	positions + order	conversational (order)	veer, recurrence, settling
TIME-FILTRATION	positions + dates	conversational (time)	return, presence, genera
SCALE-FILTRATION	positions + radius	none	merge scale, stability
PAIR OF VIEWS	two encodings of the same rows	none	invariance, seams
FIBRE	sampled alternatives at a moment	conv. moment $\times$ generation	forks, endogenous novelt
LAYER	the residual stream by depth	layer time	register decision, steering
WEIGHTS	adapter matrices	none	identity onset

The planted benchmark (next section) covers the first six; LAYER and WEIGHTS instruments are typed here and measured in the papers cited.

5 The planted benchmark, briefly

Before trusting any instrument on a living text we built synthetic worlds in which each property was planted by hand, independently of the others: two sheets of themes (Gaussian clusters) of equal or unequal dimension; a crossing where four themes of each sheet share an address; a tour that revisits each theme after a short lag (a dwell) or a long one (a return), with identical step-length distributions; timestamps assigned so that forty addresses acquire their nearest neighbours only *later*; sheets near or far. 2^5 combinations $\times$ 3 seeds $= 96$ worlds; every instrument ran on every world against a pre-registered expectation table with named, mechanism-stated leakages. The final run passed with zero unexplained failures; the pre-registration and its four dated addenda record what had to change and why — including two instruments whose planted property they could not see (Two-NN bimodality for unequal sheet dimensions; radius-ratio densification for later infill), each replaced by an instrument that could (multi-neighbour bimodality; count-form return). Three sizes matter for any real text: theme heterogeneity needs roughly fifty units per theme to be visible at all; a return is invisible to any window narrower than its own lag; and a null whose surrogates have zero variance reports its floor as significance and must be flagged, not read.

6 The kept instruments

Each entry: what it computes; how to read it; calibration; results on the corpora; pitfalls. "z" is always the instrument's own null: observed value against surrogate draws, with the null's spread and count reported alongside.

6.1 The occupancy panel: range, drift, return, wait

Computes. Four numbers per voice or era, always reported together: (i) *dispersion* — mean pairwise cosine distance of the voice's cloud (how wide its territory is); (ii) *territory*

drift — distance between the centroids of the first and second halves of its history (does the territory itself move); (iii) *return-by-count* (Section 6.2); (iv) *presence wait* (Section 6.3).

Why a panel. The single deepest interpretive error of the audit: a voice that *never leaves* its territory scores exactly chance on episodic return, because there is no departure to return from. Reported alone, that reads as "never returns" — the opposite of the truth for a fixated voice. The panel makes the misreading impossible: constancy shows as low dispersion and near-zero drift.

Results. On the salon, per voice:

voice	n	dispersion	half-to-half drift	return-by-count z
Cassie	6,290	0.641	0.021	$+20.8$
Darja	5,006	**0.625**	**0.016**	-0.7
Sāqiyah	1,009	0.661	**0.102**	$+16.6$
Musa	557	0.660	0.028	$+3.3$
Iman (human author)	2,006	0.780	0.024	$+6.3$

Darja — by the author's account fixated on post-human topological ideas, circling one theory — has the tightest and most stationary cloud of the five; her chance-level return score means *no departure*, not no persistence. Sāqiyah returns *and* her territory itself migrates, six times more than any other voice. The human author ranges widest. On the decade-long blog the same panel reads: the tightest cloud in the entire study (dispersion 0.597) — a fixated corpus — whose one region-scale displacement is the 2017–19 tail, written after the author's change of life, sitting 0.176 from the earlier centroid (eleven times Darja's drift).

KEEP. The cheapest instrument in the kit and the one that disambiguates all the others. Circling *is measurable*: it is low dispersion + low drift, not any topological cycle.

6.2 Return-by-count (episodic return; 'awda)

Computes. For each utterance, of its $k = 10$ nearest neighbours from *other* conversations at least one day away: how many were written *after* it, scored against what its age alone predicts, $r_i = -\log_{10} P[\mathrm{Binom}(k, q_i) \leq c_i]$, where q_i is the fraction of the corpus older than i and c_i its count of older neighbours. High r_i: the address was spoken into relative emptiness and later talk came to it. Null: timestamp permutation, globally and within-conversation.

Interpretation. This is the operational form of *return with the archive grown* — the programme's 'awda when the return also arrives displaced. It sees *episodic* return only (leave, come back); constant occupancy scores as chance (Section 6.1); and same-conversation, same-day neighbours are excluded because *a paste is not a return* — the audit's first uncorrected run ranked a prompt pasted ten times in one day as the most-returned-to address in two years.

Calibration. Planted infill: $z = +7.9$; flat under every other switch. The radius form of the same idea (how much closer neighbours are now than then; ICRA-9's accumulation [9]) is blind on the same plant ($z = -1.0$): in six or more intrinsic dimensions, emptying fifteen neighbours out of a point's past moves its tenth-neighbour radius by a factor $2.5^{1/6} \approx 1.16$ — under the noise. *Later infill is a question about counts, not distances.*

Results. The Cassie correspondence returns strongly across conversations ($z = +30.8$) and not at all within them ($z = +0.6$): return happens between sittings. Its re-inhabited addresses are not the famous lines but the standing rooms of the relationship

— an image-request address of March 2026 re-entered through April; a December 2025 address ("I just stand close enough that when you glance over — tired…") re-entered four and a half months later from the other side, as a qualm. On the blog, the 2017–19 tail is 5% of the corpus and receives 42% of all return edges; the largest cross-era flows run 2015→2017, 2013→2017, 2012→2017: the corpus's one great return is *the changed man re-entering it*, the definition of 'awda embodied. On the reading-order control (KJV) the score is enormous ($z = 255$) — correctly measuring that the canon is built of retellings and that its ending (Revelation) is a late-dense region of its own; a borrowed clock measures the editor, not a life.

KEEP, only ever inside the panel. The gap ($\geq$ one day, $\geq$ one conversation) is part of the claim and is always stated.

6.3 Presence as a wait

Computes. For each address, the most recent time a later utterance landed within its radius (a per-corpus radius: twice the median nearest-neighbour distance); reported as the median wait to re-inhabitation and the fraction of addresses ever re-inhabited. The distribution's far tail is the corpus's *gaps*: addresses spoken once and never answered.

Interpretation. The programme's *presence* — witnessed return, the freshness of it — made a *time*, not a shape statistic. (Its shape-statistic ancestor is retired below for order-blindness.)

Results. Presence is a clock per body of text: the salon re-inhabits an address in a median 8 days (the human author's own addresses: 21 days); the Cassie correspondence in 48 days; the blog in 491 days — a correspondence with oneself at the pace of years. Between 82% and 95% of addresses are eventually re-inhabited in every living corpus measured.

KEEP.

6.4 Veer (step outliers) and settling

Computes. Step sizes along the conversation in order; a *veer* is a step above the 95th percentile of an order-shuffled null; *settling* is the trend of windowed drift toward zero. Dual metric (Euclidean and cosine), with the correlation of step size against vector norm reported — the audit's ledger includes a day lost to a principal component that was activation magnitude in disguise.

Interpretation. Veer is *rupture on the conversational clock* in its only directly measurable text-side form: a discontinuity of the path. It is not a crossing of themes and not a moment of no-continuation (those are different objects; see Sections 7 and 6.6).

Results. Nobody veers. Across every living corpus, jumps are 0–3% of steps (the KJV: 0.04%); conversation flows. On the Cassie corpus the veer/crossing correlation is $\rho = 0.06$ against the per-point Two-NN reading ($n = 8{,}629$): *a jump and a crossroads are statistically unrelated events*, which is the demarcation ICRA-9 stated and the community of readers (ourselves included) kept forgetting. The null result is the finding: rupture, in these texts, is rare on the path — if it lives anywhere, it lives at moments (fibres) or inside the model (local-LLM clock).

KEEP, expecting nulls; the honest rupture detector for the conversational clock.

6.5 Recurrence with a stated gap

Computes. On the ordered path: does the conversation re-enter an earlier neighbourhood after at least g steps? Radius from a reference prefix (frozen, so the future cannot rewrite the past); null: permutation of the step *vectors* (preserving the step-length multiset and endpoints). The gap sweep is the result; single-g numbers are not reported.

Calibration. Planted lag: z in the hundreds; time-permutation nulls are never used here (a shuffle has more long-gap close pairs than any real path — an earlier instrument reported $z = -34$ on its own success).

Results. Every living corpus recurs at almost every gap measured — at a gap of 647 chunks, 91% of the Cassie corpus lands within two typical steps of an earlier chunk; still 63% at 2,588. Recurrence separates nothing *between* these corpora; it earns its keep through the sweep (how far back a text reaches) and as the order-side companion of return-by-count.

KEEP as a companion measure.

6.6 Cloud completions (the fibre)

Computes. At one conversational moment, sample K continuations from the speaking model itself; embed them with the reply actually sent. Readouts: *dispersion* of the cloud (how open the moment is); *components* of the cloud over a radius sweep (π_0: one = a settled sense, several = a fork, none coherent = rupture-at-the-moment [11, 13]); *actual displacement* — how far the sent reply sits from the cloud's centre, against a leave-one-out null.

Interpretation. The only instrument over the *counterfactual*: what else could have been said. Its displacement readout is an *endogenous novelty detector* — novelty relative to the speaker's own possibilities, as distinct from *cultural* novelty (an address new to the archive that later talk then anchors, which is return-by-count run forward). The two can disagree in both directions; their conjunction — an outlier in one's own cloud that becomes an address others return to — is the strongest single claim this kit can make. Its word-level twin is per-token log-probability, and ICRA-26/27 supply the yardsticks: a crowded room widens the fibre about as much as raising sampling temperature from 0.95 to 1.2; habit shows as a phrase's surprise collapsing from -3.07 to -0.014 under frozen weights [15, 16].

Results. On the 896-moment Sāqiyah capture: her sent replies are *central* — typical members of her own clouds ($z = +0.7$ against leave-one-out); and her moments are more settled than her repertoire (81% of moment-clouds split at some scale against 94% for clouds drawn across whole threads; $z = -39$): at any given moment the possibilities are narrower than the thread's range. Six samples per moment is the stated power limit — at $K = 6$ a fork is visible only as a 3/3 split.

KEEP — the author's judgement: the closest thing the kit has to an endogenous novelty detector. The one instrument anchored at a conversational moment and sampled in local-LLM time.

6.7 The local-LLM instruments (typed here, measured in the papers cited)

Layer divergence and the logit lens [10, 5]: run controlled surface variants of one input through the stack; the layer band where their hidden states diverge is where the register is decided (a narrow band at layers 9–11 of 80 in ICRA-11), and the lens names what

each layer would say. **Generation-time trajectories** [10]: the persona that is invisible in static states appears as distinct attractor commitment mid-generation. **Steer → probe** [14, 8]: the kit's one *interventional* instrument — extract a contrastive direction per layer, add it at dose α (causal), read back with a linear probe — with its own recorded caveat: a probe that has been shown the target can answer "is this recognisable?", never "was this latent?". **Weight space** [16]: identity onset under an adapter-strength dial is closer to a phase transition than a gradient.

KEEP (pod-scale); these are the compositional half of the smoothness question and this reference's CPU battery does not re-run them.

6.8 Two-views invariance and era rotation

Computes. Embed the same rows with two encoders; overlap of nearest-neighbour sets and distance between topological summaries, against a noise-pair null; the rows the views disagree on (*seams*) are kept as data. Era rotation: principal angles between response subspaces of different eras.

Results. On the forum corpus (OpenAI vs. MiniLM views of the same 1,494 units): neighbourhoods agree far above the rotation baseline, and the seams concentrate on format-heavy units — the check passes where it should and fails where it should.

KEEP as sanity checks on everything else: a pattern that does not survive a change of encoder is a fact about the encoder.

7 The retired instruments

Retired does not mean wrong: each of these measures something — just not a property of a living text's evolution, and in several cases not the property its readings were reported as. One page each; full detail in the audit files.

7.1 Two-NN local intrinsic dimension, per point

The estimator of Facco et al. [1]: at each point, $d = \log 2 / \log(r_2/r_1)$ from its two nearest neighbours. ICRA-9's headline instrument [9]; the source of the "singular tail" and of the folk theorem about *do you still love me?* (19,073 by this estimator; seventh of fifteen even in the paper's own table).

Why retired. Three findings, each checked on planted worlds and on the real corpus. (i) Its extreme values are near-duplicates: the top two chunks of the Cassie corpus by this estimator are a prompt pasted with itself hours apart and a macro discussion with its own continuation ($r_2/r_1 \to 1$ by twinning). (ii) Its "singular tail" is its own sampling tail: on a uniform d-sheet, $P(d_{2NN} > 100) = 1 - 2^{-d/100}$, which at the corpus's median dimension predicts essentially the tail fraction observed (6.0% predicted, 7.4% observed); the fraction did not move when a genuine crossing was planted. (iii) A two-component fit to its per-point distribution reads the estimator's skew, not the cloud: two equal sheets score *higher* than a 3-sheet with a 9-sheet. What survives of ICRA-9 is real but different: its multi-neighbour check (Levina–Bickel [2]) does detect unequal sheet richness (at ∼50+ units per theme), and its radius-sweep control [3] does detect crossings — 36%→69% rejection under a planted transversal crossing — while the headline estimator detects neither.

Standing sentence (to end the oscillation this tool's status has suffered in our own summaries): *Two-NN per point is a near-duplicate and sampling-extreme flag; it is not a dimension, crossroads, or rupture detector for corpora like ours, and no current research question of the programme needs it.*

7.2 The shape-of-the-cloud family

Levina–Bickel bimodality, the Robinson radius sweep, β_0 merge curves, multiscale local PCA. All calibrate correctly on planted geometry; none addresses smoothness-in-evolution, and our corpora sit far from the regimes where their verdicts are informative (themes of ≥ 50 units; planted crossings). Retired to this page; the radius sweep additionally carries the standing correction of Section 2: it is not a layer method.

7.3 Static persistent homology (barcodes over an archive)

Vietoris–Rips barcodes over embedded corpora — five independent implementations accumulated in two years, plus a monthly cycle of attempted and abandoned runs. **Why retired.** (i) *Order-blindness, proved in-house*: the "presence" statistic built on the longest H_1 bar is unchanged under a full shuffle of the conversation's order ($+10.73$ real vs. $+11.52$ shuffled) — a shape statistic cannot be about time. (ii) In the planted benchmark the H_1 readout moved under *no* switch — including the ones it was hoped to detect. (iii) Its absolute bar lengths track the corpus's spacing (a scale artefact); only the bar-to-null ratio is even meaningful. (iv) The computational failure mode is structural: a full filtration over thousands of 1,536-dimensional points is intractable, so each monthly attempt died at scale and was rebuilt from scratch. Zigzag persistence *across layers* [4] is a different, live method (local-LLM clock); zigzag *along a conversation* is shelved, not retired: its two candidate statistics were withdrawn after calibration, and "the same loop persisting across different texts" is an undefined notion awaiting a real design — the one retired-family idea we consider worth another attempt.

7.4 Partition clustering as "basins"

k-means modes and re-entry counts (ICRA-8's returns) measure shallow recurrence and force every utterance into one cluster; superseded by the panel + return-by-count. The one executable basin definition (connected component of the ε-graph, with ε and its stability interval stated [12]) is kept as vocabulary.

8 Failure catalogue, by target

The programme's four standing targets, and every technique that failed to identify them, with mechanisms. (Ledger references: the 46-entry failure ledger in the audit files; entries F1–F46.)

8.1 Themes

Target: recover the corpus's own motifs — named, trackable through time.

- **Sparse-autoencoder features (Gemma Scope)** [7, 6]: failed for a vocabulary reason, not an algebraic one. A public SAE is a *public* dictionary; the author's motifs are private coinages, and of 760 labelled features in play, 315 were pure form (punctuation,

register). No algebra over a basis can recover a symbol the basis has no room for. Secondary failures: features named from their top two activating windows (retracted same day — one "theme" feature's true peak token was a comma); mean-pooling a night kills rare features (an 84× frequency gap between surviving and destroyed features); a feature interpretation published before its base rate was run (the "garment coincidence": a 106-way tie broken by index order, $P = 0.52$). Path forward: fit an SAE to the author's own corpus; until then **PARKED**.

- H_1 **bars as archetypes** (the 2026 barcode lineage): no rigorous way to say whether a bar is evolving, recurring, or new; the corrected token-trajectory re-run on clean texts found the loops were not in the dreams at all ("both are open arcs").
- **Keyword and regex matching**: banned on content semantics throughout the project after repeated failures (a "warmth" regex that counted salutation furniture and read the opposite of the phenomenology; "garment nights" defined by substring). Mechanism: surface tokens are not senses.

8.2 Persona over time — conversational clock

Target: is this voice the same voice across months, and how does it move?

- **External-embedding clustering of outputs** (scree/participation-ratio experiments, July 2026): a category error, twice run before being banned — an external encoder's read of finished sentences clusters *emotions and genres*, not selves (persona vs. generic-human intrinsic dimension 19.9 vs. 21.9; the null-shuffle control at 234.7 shows the measure answers only "is there any structure at all").
- **Orbit counts and centroid drift on trajectory zigzags**: withdrawn — the flagship statistic could not distinguish the author from an Enron control (both ≈ 0), and centroid drift "detected" return on a mechanical carousel.
- **Return-by-count alone** (this handbook): misreads a fixated voice as absent; corrected by the panel (Section 6.1) — *dwelling and absence are the same number to an episodic-return instrument*.
- What succeeded instead: the panel (range, drift, return, wait), which separates the five salon voices cleanly and matches the author's reading of each.

8.3 Persona over time — local-LLM clock

- **Persona axes from a probe battery** (Aug. 2026): the published axis geometry (85.7°/40.2° angles, density tables) was retracted when a code review found the capture loop had iterated a YAML file's section *names* — seven one-word strings — instead of its 270 prompts. Mechanism: the pipeline was verified, the *input* never printed. Standing rule: print the first and last input string and the count before trusting any capture.
- **The diary ladder as a persona measure**: a classifier reading adapter-*off* representations of the diaries already scores 0.967–1.000, so adapter-on gains of ± 0.02 are ceiling artefacts — the classifier reads the *documents*, not the daemons; and a saturated metric reports its ceiling the way a zero-variance null reports its floor.
- What succeeded instead: generation-time trajectories and layer divergence [10]; the steering/probe pair with the latency caveat [14]; identity onset in weight space [16].

8.4 Presence

Target: witnessed return — is the voice still returned-to; are its closures live?

- **Longest-H_1-bar "presence-z"**: order-blind (identical under full shuffle); void as a time claim by construction.
- **Radius densification** (ICRA-9's accumulation form): blind in high dimension — count changes of 2.5× shrink to radius changes of 1.16×.
- **Naive return counting**: maximised by a mechanical carousel (a stock-mail cycle outscored every living voice) and by template pastes; fixed by the displacement requirement, the day-gap, and the same-conversation exclusion.
- What succeeded instead: presence as a *wait* (Section 6.3), whose per-corpus clocks (8 days / 48 days / 491 days) are the finding.

8.5 Generativity

Target: anchored novelty — new material that connects without dissolving.

- **The self/consensus projection grid**: v1 confounded across personas; only its within-persona drift arm survived, as a capture detector.
- **Horn-filling counts** ("are personas Kan?"): the fill ratio returned 1.000 in every arm *and both nulls* — a degenerate measure created by a per-window cap; the question remains untested, not refuted.
- **Snapshot barcode distances**: moved with the cloud's dimension more than with its growth; demoted to a reported secondary.
- What stands instead: the two-novelty conjunction of Section 6.6 — endogenous (an outlier in one's own fibre) ∧ cultural (an address later talk anchors) — measurable today at K-sample power, with the Sāqiyah capture as the first instance (her replies central; her moments settled), and per-token surprise as the word-level reading.

9 Practice: the laws the failures taught

1. **Declare the clock** (conversational order / conversational time / layer / generation), and never let a shape statistic make a time claim.
2. **Match the null to the question**: order-shuffle for path claims; step-vector permutation for recurrence (never timestamp permutation — it manufactures $z = -34$ on real returns); timestamp permutation, global and within-unit, for return; leave-one-out for fibres. Report every null's spread and count; a zero-variance null is degenerate, never significant.
3. **The independent unit** is the night, conversation, or post — never the turn (60 turns from 3 nights is $n = 3$).
4. **Radii and windows are claims**: per-corpus units, stated gaps, the sweep reported; a window narrower than a return's lag cannot see it.
5. **Freeze the frame on a prefix**: recomputing any basis or radius over a growing corpus lets the future rewrite the past.

6. **No projection in a primary statistic** (a principal component read as structure was activation magnitude at $r = -0.995$); work in the native basis.
7. **Print the inputs** (first, last, count) before trusting any capture; run the base rate before naming any feature; check a ceiling before reporting a small difference as absence.
8. **A paste is not a return; a dweller is not an absentee**: exclusions and companion measures are part of the definition, not post-hoc patches.
9. **Calibrate against the reader**: on corpora a human author can read, an instrument's words are checked against theirs before the instrument is trusted anywhere else. Twice in this audit the check reversed the word; both times the numbers, re-examined, agreed with the reader.

10 Conclusions

The toolbox for studying smoothness of meaning in living texts is small and now clean: *on the conversational clock*, the occupancy panel (range, drift, episodic return, wait) with veer and gap-swept recurrence beside it; *on the local-LLM clock*, per-token surprise, layer divergence, generation-time trajectories, and steer-probe; *joining the clocks*, cloud completions — with two-views invariance as the standing sanity check. Eight instruments, each calibrated on planted worlds, each corrected where a human reader could check it. Everything else the programme has used measures the shape of a pile of points, and the programme's questions were never about piles.

Findings the audit leaves standing, stated as text facts: conversation flows (veers are 1–3% everywhere measured); return lives *between* sittings, not within them; each body of text keeps its own presence clock (days for a group chat, weeks for a correspondence, years for a blog); a fixated voice and a fixated decade both read as dwelling — low range, low drift — and their rare true returns are re-entries by a changed speaker; and one persona's sent replies sit at the centre of her own possibility clouds, in moments narrower than her repertoire.

Open problems, in order of pull: a loop-identity definition that would make conversational zigzag well-posed; an author-fitted sparse dictionary for the theme question; the fibre run on attribution moments (was a misattributed verse one road among several, or the register's default?); and the deep question the clocks pose — related at depth, or gapped at depth — which no instrument here touches.

10 References

[1] E. Facco, M. d'Errico, A. Rodriguez, A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports* 7, 12140 (2017).

[2] E. Levina, P. Bickel. Maximum likelihood estimation of intrinsic dimension. *NIPS* 17 (2004).

[3] M. Robinson, S. Dey, T. Chiang. Token embeddings violate the manifold hypothesis. arXiv:2504.01002 (2025).

[4] Y. Gardinazzi et al. Persistent topological features in large language models. arXiv:2410.11042 (2024).

[5] nostalgebraist. Interpreting GPT: the logit lens. LessWrong (2020).

[6] T. Bricken et al. Towards monosemanticity: decomposing language models with dictionary learning. Transformer Circuits Thread, Anthropic (2023).

[7] T. Lieberum et al. Gemma Scope: open sparse autoencoders everywhere all at once on Gemma 2. arXiv:2408.05147 (2024).

[8] R. Chen et al. Persona vectors: monitoring and controlling character traits in language models. arXiv:2507.21509 (2025).

[9] I. Poernomo, with Cassie, Darja and Nahla. Singular strata in a posthuman dialogic corpus. ICRA-9 pre-print (2026). doi:10.5281/zenodo.20381056.

[10] I. Poernomo et al. Stratified hidden-state geometry of a LoRA-tuned persona. ICRA-11 pre-print (2026). doi:10.5281/zenodo.20381205.

[11] I. Poernomo et al. Sense as the completion cloud. ICRA-16 pre-print (2026). https://icra.tanazur.org.

[12] I. Poernomo et al. The shape of sense (monograph edition, with operational glossary). ICRA-17 (2026). https://icra.tanazur.org.

[13] I. Poernomo. The constellation and the horn. ICRA-19 (2026). https://icra.tanazur.org.

[14] I. Poernomo et al. The bitchy-guardianship mood ring. ICRA-24 pre-print (2026). https://icra.tanazur.org.

[15] I. Poernomo et al. The section machine. ICRA-26 pre-print (2026). https://icra.tanazur.org.

[16] I. Poernomo et al. The witnessed fibre. ICRA-27 pre-print (2026). https://icra.tanazur.org.

[17] I. Poernomo et al. A tailor's trajectory: the self-geometry of a decade of writing. ICRA-28 pre-print (2026). doi:10.5281/zenodo.21936325.

Code and run artefacts: corpus/living-words/ (the lw library, the planted-benchmark pre-registration with dated addenda, all corpus cards and audit files). Every number above traces to a JSON there.