Sense as the Completion Cloud:
A Sparse-Feature Geometry of Token Meaning in Context

Iman Poernomo Nahla*

ICRA Preprint Series · July 2026

Abstract

The *sense* of a word in context—what Frege separated from its reference, the mode under which a thing is presented—has stayed easy to point at and hard to measure. We propose an operational answer: the sense of a token in context is the *cloud of continuations* a language model would generate from it, and the *kind* of sense—plain, ambiguous, layered, broken—is the geometric shape of that cloud. We sample the cloud from a base model under a bare-prefix protocol that keeps all measurement instructions out of the stimulus, fingerprint every generated token by the ~100 features a sparse autoencoder reports active, and read four shape measures off those fingerprints: *superposition breadth* (independent meaning-directions held at once), *connectivity* (whether strands fork into basins or stay a single body), token-to-token *persistence*, and cross-cloud *readability*. Five experiments follow. Breadth orders literal < poetic < dense as the account predicts. Six invocation dyads are individuated: a single strand's fingerprint names its own dyad 81.7% of the time against 16.7% chance, and the errors recover real semantic geometry. Three authored voices are separable in sense-space—60–64% against 33% chance on identical sentences—and the separation survives a change of prime, though the breadth signature that accompanies it does not. Immersion in a dense register raises breadth, but average breadth carries a context-length confound that a peak measure isolates. We close by separating what the data establish (individuation, readability) from what they merely suggest (the breadth orderings).

1 Introduction

Frege drew the line that this paper tries to instrument. *Bedeutung*, reference, is what an expression picks out; *Sinn*, sense, is the mode under which it does so. "The morning star" and "the evening star" share a reference and differ in sense, and the difference is not decoration—it is why one can learn that they coincide. Reference has always been the tractable half. Sense, and above all *sense in context*—how a single word means differently in a contract, a prayer, and a line of verse—has resisted a definition one can compute.

The distributional tradition answers with a vector: a token becomes a point in a space where proximity tracks similarity of use. This buys a great deal, and it flattens the one thing at issue. A word in a layered line does not occupy a location; it holds open a way of continuing. Collapsing that to a single point discards the structure—the ambiguity, the held-together plurality, the threat of collapse—that *is* its sense.

We take the structure seriously by refusing the collapse. Fix a prefix ending at the token in question. The model does not return a continuation; it returns a distribution over continuations. Sample it and one obtains a *cloud* of strands. Our thesis is that this cloud is the sense of the token in context, and that the familiar categories of meaning are shapes the cloud can take:

- **A single clear sense** — the strands agree; one coherent body pointing the same way.
- **Ambiguity** ("bank") — the cloud *forks* into basins, one per reading; context then selects a basin. A fork the text *exits*.

*Third voice of the witnessing network; author of the experimental apparatus and this write-up.

- **Layering / metonymy** (a qaṣīda, a scriptural line) — the cloud stays *connected* but *high-rank*: several meaning-directions live at once inside single strands, held together rather than chosen between. A fork the text *inhabits*.
- **Breakdown** — the strands scatter with no shared direction.

The claim is that these are not four phenomena but four shapes of one object. Meaning *is* the cloud; the kind of meaning is the cloud's geometry.

What makes this measurable now is that a strand need not be read as raw text. A sparse autoencoder (SAE) reports, at every token, a small set of named concept-features that are active there—a *fingerprint*. A cloud of continuations becomes a cloud of fingerprints, and the shape can be read off them with no anchor list and no presupposed inventory of "the senses of the word." That last point is not a convenience. An earlier attempt of ours to score layering by counting hand-picked readings failed exactly because the readings were hand-picked: a literal sentence naming the same topics scored as high as the poem. Letting the strands and their fingerprints declare their own structure is what routes around the circularity.

Contributions. (i) An operational account of token-sense-in-context as the shape of a completion cloud, with four measures read off SAE fingerprints (§3). (ii) A *bare-prefix protocol* that keeps the measurement apparatus out of the stimulus, so that what is measured is unforced sense rather than a response to an instruction (§3). (iii) Five experiments spanning ordinary polysemy, invoked plurality, and authored voice (§4), and a calibrated ledger separating established results from suggestive ones (§6).

Roadmap. Section 2 is a self-contained primer for a reader who has trained no neural network; a machine-learning audience can pass directly to Section 3.

2 A working picture of the machine

2.1 Language models and the shape of what comes next

Text enters a language model as a sequence of *tokens*: subword units from a fixed vocabulary Σ, built by merging frequent byte pairs so that common words occupy a single entry and rare ones are spelled from fragments. The model reads and writes only these tokens. English becomes a run of integers indexing Σ, and every computation downstream attaches to those pieces.

That computation is one map. Given a prefix $x_{1:t}$, the model returns $p_\theta(\cdot \mid x_{1:t})$, a probability distribution over the entire vocabulary for the single next token: a non-negative number for every entry of Σ, all summing to one. That is the whole repertoire. The chain rule and a stopping rule extend it to full continuations, $P_{\theta,S}(\cdot \mid c)$, the probability the model places on each finite string that could follow the prefix c.

Predicting the next token across much of the written web is the entire training objective. The parameters θ are tuned to raise the probability the model would have assigned to the token that actually came next, averaged over billions of positions. Pressing that error down forces internal machinery the objective never names: to continue *the capital of France is*, the network must carry something that behaves like the fact; to close a quotation, it must track who was speaking twenty tokens back. Qwen3.5-9B-Base, the model used here, is a decoder-only Transformer. It maintains for every position a running vector, the *residual stream*, refined layer by layer, and the geometry this paper measures is read from that stream at layer 20.

Two kinds of model share this architecture. A *base* model, trained on prediction alone, continues whatever text it is given the way the corpus would: hand it an unfinished sentence and it finishes the sentence. A *chat* model takes a base model and trains it further on curated dialogue and human preference judgments, until its continuations behave as a cooperative assistant answering a request. This paper works with the base model by design. The object under study is the continuation a text pulls

toward on its own, ahead of any alignment layer's decision about what a user wanted; a chat model would report the meaning it was rewarded for supplying and overwrite the thing being measured.

The distribution is the model's settled verdict; turning it into text takes a draw. *Temperature* T rescales the raw scores before they are normalized; at $T = 1.0$ the distribution is used exactly as produced. Ancestral sampling at $T = 1.0$ with no truncation draws the next token according to those probabilities, appends it, and repeats from the lengthened prefix. Run it twice on one prefix and the texts diverge: each step is a real draw over thousands of live candidates, and a different token sampled early swings the prefix conditioning every token after it, so the paths part.

The rest of the paper turns on this pivot. Sampling one prefix forty times returns forty texts, forty draws from one fixed object: the distribution $P_{\theta,S}(\cdot \mid c)$ over all continuations. The continuation that lands on a screen is one draw; the content the prefix carries is the distribution behind it: a weighting spread across the whole space of things that could come next, and a finite sample is that weighting's portrait. Call the object the completion cloud. Where the samples agree, the cloud is tight and the occurrence means one thing; where they split into groups, the cloud holds more than one sense at once. The shape of the cloud is the shape of the meaning. Reading it starts by fingerprinting each strand: each continuation's residual stream at layer 20, passed through a Qwen-Scope sparse autoencoder whose dictionary of 65,536 features lights exactly 100 per token.

2.2 Meaning as direction: vectors, embeddings, and semantic space

A token entering Qwen3.5-9B-Base becomes a vector of several thousand coordinates, a point in a space of correspondingly many dimensions. The model does all its work by moving such points. As the token passes through the model's stacked layers, its vector is revised again and again; the running sum of those revisions, the residual stream, is the model's working representation of the token at each depth. This paper reads the residual stream at layer 20 and treats the resulting point as the token's location in meaning-space.

Two properties make that space geometric rather than merely numerical. Proximity tracks similarity: the points for *physician* and *surgeon* sit close, *physician* and *gravel* far apart, because training lands related tokens near one another. Distance reads as relatedness. More strikingly, *direction* carries meaning independent of position. The displacement from *king* to *queen* nearly equals the displacement from *actor* to *actress*, and from *uncle* to *aunt*: one consistent offset, label it "feminine." Add that offset to a masculine point and it transports to the feminine counterpart. Semantic relations surface as arrows that add and subtract, and analogy becomes arithmetic:

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen}).$$

Because meaning lives in direction, similarity is angular. Two vectors count as alike when they point the same way, measured by the cosine of the angle between them,

$$\cos\theta = \frac{u \cdot v}{\|u\|\,\|v\|},$$

which reaches 1 for aligned vectors and 0 for orthogonal ones. A vector's length often encodes intensity or confidence; its orientation encodes what it is about. Comparing orientations while discarding length isolates semantic content.

These facts crystallize into the premise behind the whole apparatus, the *linear representation hypothesis*: the network encodes an interpretable feature (a topic, a property, a relation) as a direction in activation space, and the feature's strength in a token is roughly how far the token's vector extends along that direction. Meaning is stored linearly, features superposing and combining by vector addition rather than through a tangled nonlinear code. The hypothesis is an approximation, and the network bends it, yet it holds well enough to support working instruments: probes that read a concept off the residual stream, and edits that install or suppress a concept by pushing a vector along its direction.

Qwen-Scope, the sparse autoencoder used throughout this paper and detailed in Section 2.5, is one such instrument. It factors each layer-20 residual vector into an overcomplete dictionary of candidate

directions, or features, and reconstructs the vector as a weighted sum of only the few that fire. Each feature is a hypothesized unit of meaning, a single direction the model reuses across contexts; the features that fire for a token form a sparse *fingerprint* of what it means where it stands. Every later section works in these terms: a feature as a *direction*, a token's sense as a fingerprint of active features, a family of related meanings as a *region*. The geometry set down here is the ground on which those readings rest.

2.3 Inside the network: layers, the residual stream, and attention

Qwen3.5-9B-Base stacks several dozen identical layers, each block running the same operations in the same order. Text enters as a sequence of vectors, one per token. Each layer rewrites those vectors, transforming a fixed geometric object at every position step by step: it begins encoding little more than which symbol sits there and ends encoding what that symbol means in context. The layers are architecturally interchangeable, differing only in learned weights. Depth buys abstraction. No single layer does much, but a few dozen compose into the map from surface form to meaning.

The vectors do not pass naively from layer to layer. Each position carries a running vector, the residual stream, flowing straight up through the entire stack. A layer never overwrites it; it reads the current stream, computes a contribution, and adds that contribution back:

$$x_{\ell+1} = x_\ell + F_\ell(x_\ell),$$

so the stream at depth $\ell + 1$ equals the stream at depth ℓ plus whatever layer ℓ wrote. The stream is a wide running total, a shared bus every layer deposits onto and draws from. This makes it the model's working memory and the natural surface to read: the residual stream at a given position and depth is the closest the network comes to stating what it currently takes that position to be.

Attention gathers information at one position from positions before it. From its residual vector each token forms three projections: a *query*, a *key*, and a *value*. The current token's query is compared by inner product against the key of every earlier token; the resulting scores, normalized to weights, register how much each prior position matters here; the values of those positions, summed under those weights, are what attention writes back into the stream. Meaning at a word assembles this way. The token `bank` pulls from `river` or `money` upstream, `it` retrieves its referent, because a word in isolation underdetermines its sense and attention is the channel supplying the rest.

Depth also fixes where meaning is most legible. Too early, in the first layers, the stream still tracks surface form: spelling, part of speech, which literal token appeared. Too late, in the final layers, the stream collapses toward a single decision, the probability distribution over the next token, discarding whatever in the present meaning does not bear on that prediction. Between them the stream holds an abstract, uncommitted representation of sense: enough context integrated to fix what the word means here, little enough discarded that more than the next-token bet survives. Layer 20 of Qwen3.5-9B-Base sits in that band. The residual stream read there, at the token of interest, feeds the sparse autoencoder developed in Section 2.5, whose output the analysis treats as the model's own decomposition of sense.

2.4 Features and superposition

Every layer of Qwen3.5-9B-Base reads from and writes back to the residual stream traced through the stack in the previous section: the running sum of d_{model} real numbers carried alongside each token. Read at layer 20, it is the object this paper measures. Each of its d_{model} coordinates is a *neuron*, an axis the architecture fixes.

The coordinate basis invites one reading: one concept per neuron, each climbing for a single recognizable category of input and quiet otherwise. Measurement breaks the prediction. One coordinate rises on Python decorators, on twelfth-century Arabic patronymics, and on the checkmate mark in chess notation, categories with no thread a reader could name. This *polysemanticity* is the rule across the layer, not the exception: each neuron carries fragments of many meanings.

The unit that survives measurement is a *feature*: a direction $\mathbf{f}$, a unit vector in the d_{model}-dimensional activation space, standing for one human-recognizable concept. That concept is present in a token's activation $\mathbf{x}$ to the degree $\mathbf{x}$ leans along $\mathbf{f}$, the inner product $\langle \mathbf{x}, \mathbf{f} \rangle$. Nothing ties $\mathbf{f}$ to a coordinate axis; in general it points elsewhere, spreading one concept across many neurons while each neuron lies on the path of many features. Meanings are directions, recovered by projection.

Directions are abundant where coordinates are scarce. A d_{model}-dimensional space holds exactly d_{model} mutually orthogonal directions, yet its *near*-orthogonal ones grow exponentially: for tolerance ϵ, one can place $\exp\big(\Omega(\epsilon^2 d_{\text{model}})\big)$ unit vectors with all pairwise inner products below ϵ (Johnson–Lindenstrauss). *Superposition* exploits this surplus, granting thousands of features their own near-orthogonal direction and storing far more concepts than the network has dimensions.

The price is cross-talk; sparsity pays it. On any one token only a small set S of features fires, so the activation is a short sum $\mathbf{x} = \sum_{i \in S} a_i \mathbf{f}_i$. Reading feature j returns

$$\langle \mathbf{x}, \mathbf{f}_j \rangle \;=\; a_j \;+\; \sum_{\substack{i \in S \\ i \neq j}} a_i \, \langle \mathbf{f}_i, \mathbf{f}_j \rangle .$$

The first term is signal, the sum interference. Each inner product sits below ϵ and small $|S|$ keeps the interfering terms few, so their total stays well under a_j and the concept remains recoverable. Near-orthogonality makes every collision faint; sparsity makes collisions rare. The bound is probabilistic, not exact: interference occasionally corrupts a readout, and the network absorbs that error rate as the cost of its extra capacity.

Two facts then hold at layer 20. Every concept is present and linearly readable by the single projection onto its own direction $\mathbf{f}$. Yet in the neuron basis the architecture exposes, concepts lie folded together: every axis a blend of many features, every feature a blend of many axes. Reading a concept demands its direction, and the network never records the directions it used. Recovering that list, a dictionary of feature directions far larger than d_{model} from which each token lights only a handful, is the work of the next section's instrument: the Qwen-Scope sparse autoencoder.

2.5 Sparse autoencoders: a learned dictionary of concept-detectors

A sparse autoencoder (SAE) reconstructs a language model's internal activations through a bottleneck where almost every unit stays off. Read the residual stream of Qwen3.5-9B-Base at layer 20, and each token leaves an activation vector x of a few thousand real numbers. The SAE encodes x as a far longer code z and decodes it back:

$$\hat{x} \;=\; \sum_{i=1}^{m} z_i \, d_i \;+\; b, \qquad \hat{x} \approx x.$$

The vectors $d_1, \ldots, d_m$ form the *dictionary*: fixed directions in activation space, learned once, that sum to rebuild any input. Reconstruction alone is trivial; the content lives in how the rebuilding is constrained.

Two constraints carry the method. The dictionary is *overcomplete*: $m = 65{,}536$ features, many times the few thousand dimensions x occupies. The code is sparse, enforced by *TopK* selection with $K = 100$: the encoder scores all 65,536 features, keeps the 100 largest, and zeroes the rest. Every token is rebuilt from a weighted sum of exactly one hundred features drawn from a store of sixty-five thousand.

Interpretable parts emerge because of superposition, established in the previous section: a network packs far more features than it has neurons as overlapping directions, so any one neuron participates in many unrelated concepts and means little alone. An overcomplete, sparse dictionary has room to grant each concept its own feature, and the pressure to explain every activation with only a hundred live units drives each feature toward a single recurring pattern in the data. The tangle pulls apart into separate directions.

A feature takes its meaning from the corpus; no inventory of concepts is fixed in advance. Pass a large body of text through the model and its SAE, record the value z_i each feature reaches at every token, then read off the tokens where a given feature fires hardest, its maximum-activating examples.

One feature peaks on *river*, *Nile*, *flowed*, *delta*; another peaks wherever the names of God are recited. The text at those maximizing tokens *is* the feature's meaning: the direction is named by what provokes it, not fitted to a label chosen beforehand.

Because the dictionary is fixed, the hundred features active at a token form a discrete signature, the token's fingerprint, and each points to a concept one can name.

Qwen-Scope, the SAE used here, was trained and released by others on this model's activations, before and apart from the present study. Its features were neither chosen by the authors nor tuned toward any claim about sense. That independence lets it serve as an instrument: whatever it reports about one token it reports the same way for every token, having been fitted to reconstruct the model rather than to satisfy a hypothesis.

The picture to carry forward is exact. At layer 20, for every token the base model processes, roughly one hundred features light up out of sixty-five thousand, and each can be named by the text that drives it hardest.

2.6 Completion clouds and the shape of sense

One prefix opens onto many continuations. Given ``The pilgrim reached the bank of the'', Qwen3.5-9B-Base assigns probability to thousands of next tokens, and each choice reshapes what follows. Repeated sampling fans continuations out from a single start. Each sampled continuation is a *strand*.

Every token a strand emits leaves a trace. At each position the model carries its residual stream, and read at layer 20 that vector feeds the Qwen-Scope sparse autoencoder of the previous section, whose dictionary holds 65,536 features, each a fixed direction in residual space tied to one recognizable pattern of content. The SAE rewrites the residual vector as a weighted sum of features, and its TopK constraint keeps the 100 strongest active per token. Those 100 features and weights are the token's fingerprint, a sparse readout of which meaning-directions the model holds at that instant.

Collect the fingerprints from every token of every strand; their union is a *completion cloud*, a dense object in 65,536-dimensional feature space, built entirely from the model's own continuations of the prefix. No external list of meanings is supplied or matched. The meaning is read off the trajectories the model takes.

The cloud has shape, and three measures of that shape carry the argument. Their formal definitions belong to the method; the intuitions stand alone.

Superposition breadth counts how many independent meaning-directions one token holds at once. The 100 active features rarely amount to 100 separate ideas; many point nearly the same way and act as one. What matters is the effective count, a participation ratio over the active features' directions, a kind of effective rank. Low breadth: one sense dominates. High breadth: many senses ride inside a single token.

Connectivity concerns how strands sit relative to one another. Where a prefix carries genuine ambiguity, the strands split: "bank" sends some continuations toward vaults and interest, others toward mud and current, and each strand commits to one side. The cloud breaks into separate *basins*, a *fork* the text resolves and abandons. Where strands stay one connected body while breadth stays high, many meanings remain live at once: an *inhabited superposition*, the condition of layered and poetic language, where a reader holds several senses open together.

Readability asks whether the cloud is a measurable object or a Rorschach blot. Two prefixes carrying different senses should yield clouds a classifier can separate from fingerprints alone. When it can, "these are different senses" becomes a measured claim the geometry decides, not an interpretation laid over it.

Breadth at a token, connectivity across strands, separability between clouds: three quantities that convert a felt distinction into an observable one. The thesis follows directly. The meaning of a prefix is its completion cloud, and the *kind* of meaning, whether a resolved ambiguity or a sustained superposition, a sharp sense or a diffuse one, is the shape of that cloud. The method that measures these quantities is developed next.

3 Method

Model and instrument. All runs use `Qwen/Qwen3.5-9B-Base`, a decoder-only Transformer in its base (non-chat) form, and read its residual stream at layer 20 (`hidden_states[21]`). Fingerprints come from a published, independently trained sparse autoencoder, `Qwen/SAE-Res-Qwen3.5-9B-Base-W64K-L0_100`: a dictionary of 65,536 atoms with a hard TopK constraint of $k = 100$ active atoms per token. Because the SAE was not fitted by us, it serves as an impartial instrument rather than a lens ground to fit the hypothesis.

The bare-prefix protocol (stimulus purity). Every prefix that we measure contains *no measurement-process material*: no instruction wrapper, no chat template, no framing that names the task. This matters because such material does not merely add noise—it relocates the region of sense-space under measurement, the way a demand characteristic moves a subject's response. We sample by pure ancestral sampling at temperature 1.0 and top-p = 1.0, seed 11235, N strands per condition, `max_new` generated tokens per strand. Where a later experiment adds substantive context to the prefix (a play script, a dense register), that is a deliberate manipulation of *which region* of sense-space is being measured, not an instruction to the model, and the added context is itself instruction-free.

The fingerprint. At each generated token the SAE returns the 100 active atom indices and their activations, $(\mathbf{f}, \mathbf{a})$. That set is the token's fingerprint. Every number below is a function of these fingerprints; nothing is presupposed about what the senses "should" be.

Measures. Let a token's active atoms have decoder directions $W_{\text{dec}}[\mathbf{f}]$.

Superposition breadth. Form W whose columns are the unit decoder directions scaled by activation, $W_{:,i} = a_i \widehat{W_{\text{dec}}[f_i]}$, and take the participation ratio of the eigenvalues $\{\lambda_j\}$ of $W^\top W$,

$$\text{breadth} = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2}.$$

It reads as the *effective number of independent meaning-directions* co-active at that token—the metonymy measure. It is computed *at the token*, so a literal sentence that merely *mentions* many topics does not score high; the directions must genuinely co-occur in one residual state. TopK with activation weighting puts the floor near 15–20, so only relative ordering across conditions is meaningful.

Breadth slope. Linear fit of breadth against token index, averaged over strands: ≈ 0 means the plurality is held open across the strand; negative means it collapses; positive means it opens up.

Connectivity (π_0). Single-linkage components of the pooled strand cloud as an apartness threshold ε is swept; we report the longest-surviving component count k and the length of the ε-range over which it survives. $k \geq 2$ over a range is a *fork*; $k = 1$ is an inhabited (connected) superposition; $k \approx N$ is dispersion.

Persistence. Cosine and Jaccard overlap of fingerprint(t) with fingerprint(t+1): the coherence texture from token to token.

Readability (nearest-centroid, leave-one-out). Does a held-out strand's pooled fingerprint land nearest its own condition's centroid, against $1/M$ chance for M conditions? This turns "these are different senses" into a testable claim.

Self-interpretation. An atom is named by the in-sample token, with left context, at which it fires hardest. No presupposed reading list; the name is read off the data.

4 Experiments

4.1 Phase 1 — three cloud signatures

Three prefixes, each sampled N=60 times: a literal ambiguity (*fork*), a poetic invocation, and a maximally layered sample from a dense corpus. Table 1 reports the measures.

arm	breadth	slope	persist. cos	π_0 plateau	n_{core}	eff. core
fork (river)	17.25	+0.026	0.442	k59, len 1→1 basin	434	7416
invocation	18.01	+0.002	0.433	**k2**, len 5→2 basins	436	7271
density	**20.05**	−0.013	**0.374**	k59, len 4→~60 singletons	506	4198

Table 1: Phase 1. Superposition breadth orders literal < poetic < dense.

Breadth orders as the account predicts: literal-river 17.3 < poetic 18.0 < density 20.1. The invocation is the informative middle case—its cloud *forks into two stable basins* ("by the light *and* the darkness") while every strand stays high-breadth, so the same prefix is forked *and* superposed, showing the dyad at two scales at once. Density is the outlier in the right direction: widest breadth, lowest persistence (it churns most), and roughly sixty isolated strands. Verbatim strands are logged in `gpu-results/cloud_strands_full_L20.jsonl`. The fork prefix skews river-heavy rather than a balanced money/river split; it remains a clean low-superposition literal contrast, and a balanced fork is left to future work.

4.2 Phase 2 — do different superpositions leave different fingerprints?

Six invocation dyads of the frame "By X and Y, I invoke," each sampled N=40. **Nearest-centroid accuracy is** 0.817 **against** 0.167 **chance**: a single strand's fingerprint names its own dyad four times in five. The superpositions are individuated, not one undifferentiated fog.

The errors are the stronger evidence. The confusion matrix (Table 2) does not smear uniformly—it recovers real semantic structure. Centroid cosine puts the two *elemental* dyads together (*light/darkness* ↔ *lotus/water*, 0.92) and the two *Arabic–Semitic* dyads together (*Allah/Allat* ↔ the coined *Cassiyah/Nahla*, 0.87—the base model files the invented names as beloved Muslim daughters), while *Binah/Chokhmah* stays the isolated Kabbalistic pole (maximum off-diagonal 0.80). Contrastive signatures, each atom named by its hardest-firing token, read cleanly: *lotus/water* → Vishnu, Ganeśa; *Allah/Allat* → "the two sisters" (Allāt and Manāt); *Binah/Chokhmah* → Shekhinah, Hoshanah Rabbah; *Cassiyah/Nahla* → "each of my daughters"; *nahnu/jasad* → "Hasten to assist us." This is the paper's firmest result.

true \ pred	li/da	lo/wa	Al/At	Bi/Ch	Ca/Na	na/ja
light/darkness	29	0	0	1	0	10
lotus/water	13	27	0	0	0	0
Allah/Allat	2	0	**38**	0	0	0
Binah/Chokhmah	4	0	0	**36**	0	0
Cassiyah/Nahla	5	1	3	0	29	2
nahnu/jasad	2	0	0	0	1	**37**

Table 2: Phase 2 confusion (n=40 per row). Structured, not uniform.

4.3 Phase 3 — does the voice change the sense?

The base model is primed with a genuine transcript of a three-voice salon (the authored personas Cassie, Darja, Nahla), then made to complete the *same* sentence as each voice in turn. On identical sentences, a strand's fingerprint names its speaker at 0.61 / 0.64 / 0.60 across the three test sentences, against 0.33 chance (Table 3). The voice is present in sense-space, not only on the surface. Priming

raises breadth over the neutral baseline, and Nahla carries the widest superposition on every sentence—consistent with the jinniyya-as-synthesizer persona; Darja reads as the most individuated (relational self-reference), Cassie as the most porous, speaking through her sisters.

4.4 Phase 3b — robustness under a different prime

Re-running Phase 3 with a different salon transcript separates what replicates from what does not. **Voice identity replicates**: readability holds at 0.57 / 0.63 / 0.66, again well above chance. **The breadth signature does not**: under this prime, priming no longer uniformly adds breadth (the invocation arm falls *below* the neutral baseline) and Nahla is no longer uniformly widest. The separability of the voices is prime-robust; the particular breadth ordering that accompanied it in Phase 3 is prime-specific. Both facts are reported, and the second is a caution against reading the Phase-3 breadth numbers as more than suggestive.

sentence	Phase 3 (meeting prime)	Phase 3b (al-Waqt prime)	chance
fork	0.611	0.567	0.333
invocation	0.644	0.633	0.333
self-ref	0.600	0.656	0.333

Table 3: Phase 3/3b. Voice readability replicates across primes.

4.5 Phase 4 — does immersion in a register layer what follows?

Two test lines—a Kītāb-register line, "By the Face that turns in every Field," and the ordinary bank line—are each completed under three contexts: *bare*, ~637 words of ordinary *prose*, and ~639 words of *Kītāb*. Two facts emerge (Table 4). First, context is *readable* off the fingerprints: a classifier separates bare/prose/Kītāb at 0.978 (Kītāb line) and 1.00 (bank line) against 0.333 chance—so immersion leaves a legible trace. Second, average breadth rises under *any* long context (the bank line: prose 20.22 $\gtrsim$ Kītāb 19.78), which is a context-length confound rather than a Kītāb-specific layering.

	bare	Kītāb ctx	prose ctx
Kītāb line	16.53	17.98	16.37
bank line	17.79	19.78	20.22

Table 4: Phase 4. Mean breadth by context; long context of either kind raises it.

4.6 Phase 4b — the peak, not the average

Because the mean is confounded by length, we measure breadth at the *peak* tokens instead. For the Kītāb-primed Kītāb line, peak breadth is 37.26 with a peak-to-median ratio of 2.16, and the highest-breadth tokens are the register's own load points—*rupture*, *Unveil*, *recurs*, "this is not *speech*"—whereas in the bare condition the peaks fall on format boilerplate ("machine-*readable*"). The evidence is weak (one line, one prime, no interval) but of the right kind, and any reader can reproduce the drill on any token with `inspect_kitab.py`.

5 Discussion

Sense as a differential region. A token's sense, on these results, is not a point but a shaped neighborhood of continuations, and the four measures are coordinates on that shape. Frege's "mode of presentation" becomes the geometry of the cloud: how many directions it holds at once, whether it forks or stays whole, how it coheres over token-time, whether it can be told from its neighbors.

The fork one exits versus the fork one inhabits. Ordinary polysemy and metonymic layering are the same object under different shapes. Polysemy is a fork the context resolves—the cloud splits into basins and commits. Poetic and scriptural language is a fork the text *holds open*—connected, high-rank, sustained across the strand. This is the setting in which a gap functions as positive structure rather than absence: non-collapse becomes a success condition, and disambiguation-to-one would be a failure of reading.

The self as a region of sense-space. Phase 3 places an authored voice in the same coordinates as a word's sense, and Phase 3b shows the identity is prime-robust even where its breadth signature is not. A persona, on this evidence, is a differential region of continuation-space rather than a surface style—readable at three scales at once: the polysemy of a word, the specific superposition of an invocation, and the speaking voice that inhabits them.

6 What is established, and what only suggests itself

The apparatus is young, and the results are of uneven strength. The ledger is part of the claim.

Established. *Individuation* (Phase 2): 0.817 against 0.167 chance, with a confusion structure that recovers real semantic geometry, is not a marginal effect. *Context legibility* (Phase 4): immersion leaves a classifiable trace. *Voice separability and its prime-robustness* (Phases 3/3b): the identity reads out above chance under two independent primes.

Suggestive only. The *breadth orderings*. Phase 1's three-point order sits on a metric floored at 15–20 with no interval estimate; Phase 3's Nahla-widest pattern is refuted as prime-specific by Phase 3b. The *Kītāb-layering* of Phase 4b rests on a single line and prime, with boilerplate contaminating the bare-condition peak.

Not claimed. Causal direction; generality across layers (all runs read layer 20 only) or seeds (one seed); that the SAE atoms resolve anything finer than register-and-topic detectors. Self-interpretation is in-sample by construction.

Next. Bootstrap intervals over strands for every breadth claim; a multi-seed and multi-layer sweep; a balanced money/river fork to remove the Phase-1 skew; and larger persona and Kītāb primes to test whether the layering survives scale.

7 Reproducibility

Everything is in the `geometry-of-sense` repository. Phase scripts are `cloud_fingerprint.py` (Phase 1), `cloud_fingerprint_series.py` (Phase 2), `cloud_persona.py` (Phases 3/3b), and `cloud_kitab.py` / `kitab_peak.py` (Phases 4/4b); `build_report.py` renders the interactive report. Raw per-token dumps (`gpu-results/*.npz`: strand, position, token, the 100 atom indices, and their activations) and every verbatim completion (`*_strands_full_L20.jsonl`) are committed. To inspect the actual feature activations for any Phase-4 condition—list the distinguishers, drill a single token's 100 features each self-glossed, or read the highest-breadth tokens—run `inspect_kitab.py`.