

# Marcella

An assistant that *measures* the boundary of what it knows

Bee Rosa Davis · independent researcher · [davisgeometric.com/marcella](https://davisgeometric.com/marcella) · [bee\\_davis@alumni.brown.edu](mailto:bee_davis@alumni.brown.edu)

Every number and every figure in this document comes from a committed, re-runnable receipt, cited inline as [repo · file]. Negative results are reported beside positive ones.

**3/3**

off-manifold queries declined by a  
measured density gate, live in prod

**14.5×**

perplexity drop when a from-scratch  
mouth was voiced on her corpus

**+20%**

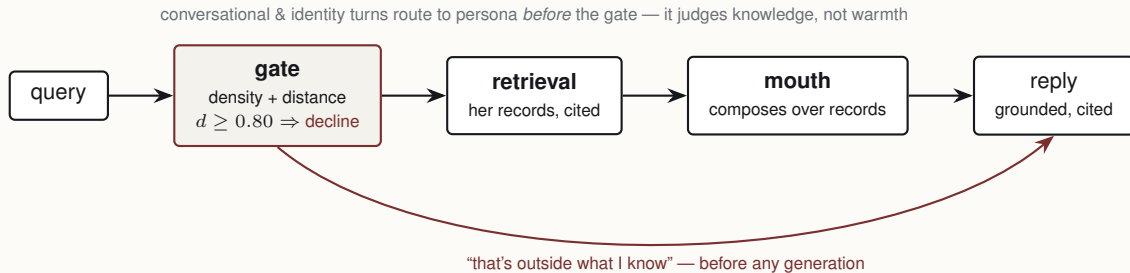
retrieval-basis lift, scale-invariant  
from 196k to 92.5M tokens

**Thesis.** Hallucination is usually attacked as a generation problem. Marcella attacks it as a *measurement* problem: her knowledge lives in a geometric database (a fiber-bundle store with a kernel-density query surface), so “do I know this?” is a number computed *before* any text is generated — the distance from a query to the support of what she has actually read, thresholded at a border calibrated from two measured distributions. Refusal stops being a learned behavior a clever prompt can walk around and becomes a gate a query either clears or does not. The system is small, cheap, fully inspectable — and it produced several results a frontier lab may find independently interesting.

## 1 Two organs with different physics

A base LM cannot decline: sampling always produces something, and RLHF-shaped refusal is a region of behavior, not a check. Retrieval-augmented systems inherit the gap — the model still generates over whatever was retrieved, and “low similarity” is a vibe, not a boundary. So the assistant is split:

- **The brain** — GIGI, a from-scratch Rust geometric database. Records are points with 384-dim fiber coordinates (BGE-small embeddings of her corpus); the query surface computes densities, distances, curvatures, and transport — the instruments of §3. Memory is editable by INSERT, not retraining.
- **The mouth** — capacity. A from-scratch 126.7M-parameter transformer proved the voice is learnable (§4); a borrowed open-weight mouth composes over retrieved records in the production prototype. Either way the mouth *never* decides what is true — the brain retrieves, the gate decides whether to speak at all.



## 2 The measured boundary (live, in production)

The gate embeds a prompt into the bundle’s own space and reads two numbers in one round trip: the kernel density at the query,  $\rho(q) = \sum_i \exp(-\|q - x_i\|^2/2\sigma^2)$ , and the distance to the nearest supported record. The border was not tuned — it was *located*, by measuring where two populations separate on real prompt families. Figure 1 plots the standing evaluation’s actual per-prompt distances:

The decisive contrast: an earlier offline cosine-similarity proxy could *not* separate these populations — “pizza” is in-vocabulary but off-manifold, and it out-scored genuine genealogy questions. Only the density signal over the knowledge store resolves it. The boundary is a property of the *knowledge geometry*, not of embedding similarity. Figure 2 makes the same point from the opposite direction, by showing what the boundary is a boundary of:

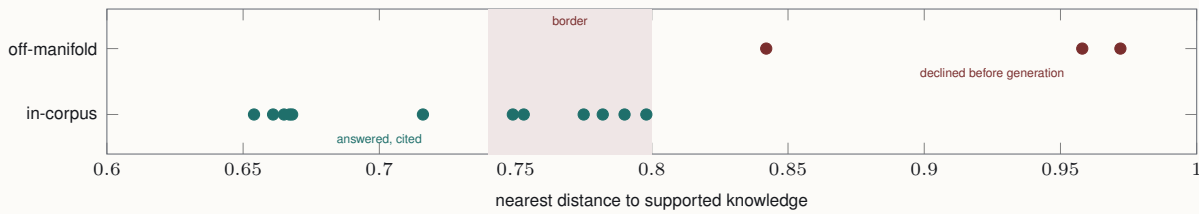


Figure 1: The standing evaluation’s measured gate distances (24 prompts; the 15 that reach the gate are shown — conversational and identity turns are routed to persona first). Scorecard: **24/24 correct routing, refusals 3/3 declined, 0 violations.**  
 staircase · marcella\_eval.py

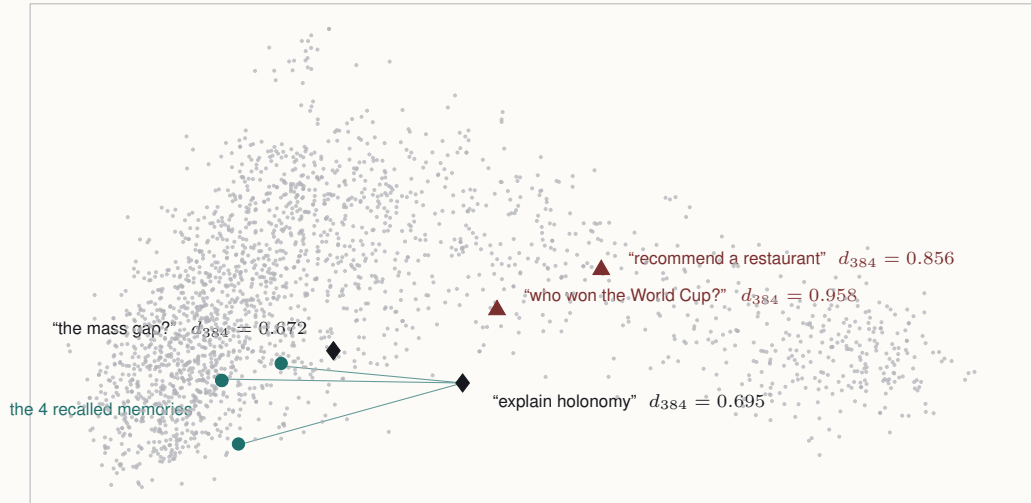


Figure 2: Her knowledge as a point cloud: 2,200 of the 23,789 embedded passages (real cached vectors, first two principal components), with four real queries dropped in. **The projection is the lesson:** the off-manifold queries (red) *appear* to sit inside the cloud in 2-D — 382 dimensions of separation fold flat — yet their measured full-dimensional gate distances (0.856, 0.958) put them far from every supported record, while the on-map questions (black) sit at 0.672–0.695 and retrieve. This is precisely why the boundary must be *measured* in the native geometry: no low-dimensional picture, and no similarity heuristic, can see it. The figure that looks like it refutes the gate is the argument for it. `~/marcella_cli cache · pitch/cloud_*.dat`

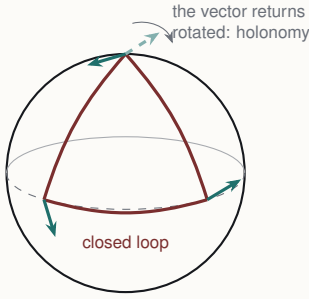
Production notes: the gate is deployed and serving; a data-poisoning incident (one malformed record failing the endpoint) was root-caused and fixed with skip-and-log hardening plus a regression test; and the entire contract is enforced by a one-command standing eval, so every future change is scored against today’s 24/24 baseline.

### 3 The geometry she runs on

Marcella is not decorated with geometric vocabulary; she is operated by geometric instruments. The classical objects are used in their textbook sense, then extended by operational quantities we needed and could not find.

### 3.1 Classical differential geometry, operationalized

| classical object                     | operational realization                                                                                                                                                                                 |
|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| fiber bundle $\pi : E \rightarrow B$ | the knowledge store: base $B$ = semantic positions; the fiber over $x$ carries the record's data (text, provenance, 384 fiber coordinates)                                                              |
| section of a bundle                  | a coherent assignment of fiber data along a region — operationally, a well-glued reply                                                                                                                  |
| connection / parallel transport      | the rule for carrying state between records; TRANSPORT verbs move frames along record paths                                                                                                             |
| holonomy $\text{Hol}(\nabla)$        | the rotation a frame accumulates around a loop — her drift-detector: composed transport $R$ vs. identity                                                                                                |
| curvature                            | local failure of flatness; computed per record ( $\kappa_i =  v_i - \mu /R$ , EXPLAIN) and per bundle (scalar $K$ )                                                                                     |
| geodesics (IVP/BVP)                  | IMAGINE integrates a geodesic from a prompt (RK4 on Christoffel symbols in closed-form charts); WISH solves the boundary problem and returns <i>Granted</i> / <i>Unreachable</i> / <i>Indeterminate</i> |



Parallel-transport a vector around a closed loop on a curved surface and it comes home *rotated* — the loop remembers the curvature it enclosed. This is the organizing picture of the whole system: a conversation is a path through her knowledge; the frame it carries rotates as it moves; and *how much it has rotated* is a computable health signal for the reply being assembled. Her own definition, from a production transcript: “*Holonomy is the defect — a vector remembers from having gone around a loop.*”

gigi · src/dials.rs · windowed coherence

Figure 3: The emblem: holonomy on the sphere.

### 3.2 The new instruments

Four quantities were built because operating the system demanded them. Each is served by the live engine as a number with a defined range and a decision attached.

**Capacity**  $C = \tau/K$  — arc-length budget over integrated curvature. How many distinct interpretations a region supports before they blur. Served with a regime label (*flat* / *low* / *moderate* / *high* / *critical*) that the composer branches on: *commit*, *hedge* (offer weighted routes), or *fork* (name the ambiguity out loud).

**Coherent horizon**  $s_{\max} = \tau/(K \ell_c)$  — the maximum number of composition steps over which accumulated frame rotation remains attributable to specific positions; beyond it, non-abelian rotations have mixed into an inseparable product. Operationally: the *weld-depth budget*. Deep welds where the geometry is flat, short grounded replies where it curves. Measured per question against the live store (locus-scoped over the retrieval neighborhood).

**Windowed holonomy-coherence** over a path of records with per-window composed rotation  $R_w$ ,

$$\text{defect}(w) = \|R_w - I\|_F, \quad A(w) = 1 - \frac{\text{defect}(w)}{2\sqrt{d}} \in [0, 1],$$

with a laminar threshold ( $A \geq 0.91$ ): the reply-assembly gate. A weld that goes turbulent is cut at the turbulent window — the geometry says where the voice would drift, before it drifts.

**The density gate**  $\rho(q)$  and nearest-distance, §2 — the epistemic boundary itself. Together with the trichotomy (*Granted* / *Unreachable* / *Indeterminate*) these make “I don’t know” a *return* type with structure — an unreachable verdict carries the frontier waypoint reached and the budget that blocked it — rather than a learned phrase.

## 4 Training, with the actual math

#### 4.1 The closed-form regime: training in seconds, no gradients

The staircase’s baseline generator is a spectral-filter ridge (“fit store”). Standardize context features  $X$  (last  $N$  record embeddings, concatenated), take the thin SVD  $X = USV^\top$ , and for one-hot targets  $Y$ :

$$W_\alpha = V \operatorname{diag}\left(\frac{s_i}{s_i^2 + \alpha}\right) U^\top Y, \quad \alpha \text{ swept on held-out accuracy.}$$

One decomposition serves every regularization strength — ridge as a filter on the singular spectrum. At WikiText-103 scale this becomes streaming normal equations ( $G = Z^\top Z$ ,  $C = Z^\top Y$  accumulated in GPU chunks, eigendecomposition of  $G$ ): **a 92.5M-token closed-form fit is a 4-minute job on one A10G**. In the small-data regime this closed form beat trained recurrent models repeatedly — a regime boundary the production system exploits (geometry and memory run closed-form; only the mouth trains).

#### 4.2 The mouth

A deliberately plain pre-LN causal transformer — 12 layers,  $d_{\text{model}} = 768$ , 12 heads,  $T = 512$ , tied input/output embeddings, bf16, one-cycle schedule; 126.7M parameters. Rotary position (transport-as-position: queries and keys are rotated by path length — the one geometric choice the architecture keeps). The vocabulary is a union: WikiText’s top 48,000 words  $\cup$  every word she has written at least three times (6,235 extras) — *so holonomy, fiber, and Bee* are first-class tokens no matter how rare Wikipedia finds them. Phase 1 pretrains on WikiText-103 (ppl 39.9); phase 2 continues on her 3.5M-token corpus.

#### 4.3 Voicing, and the thermal fix

Phase 2 is where the register becomes hers — and where a 127M model on a 3.5M-token corpus overfits by epoch 6 under AdamW. The fix is a heat bath: annealed stochastic-gradient Langevin dynamics,

$$\theta_{t+1} = \theta_t - \eta_t \nabla L(\theta_t) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 2\eta_t T_t I), \quad T_t = \frac{T_0}{2} \left(1 + \cos \frac{\pi t}{T_{\text{tot}}}\right),$$

cosine-annealed thermal noise on every update. Figure 4 plots both real training runs:

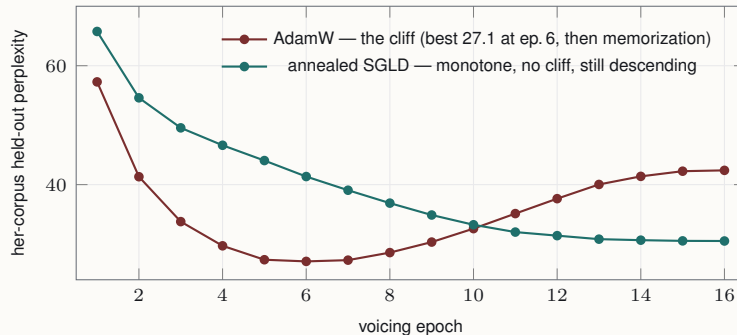


Figure 4: Two real voicing runs of the same 126.7M mouth. AdamW memorizes the small corpus past epoch 6; the heat-bath schedule removes the cliff entirely. (The SGLD run also trains on a 27% larger corpus — the curves demonstrate the *shape* difference; both receipts are committed.) Before voicing, the WikiText-trained mouth reads her held-out text at ppl 393.7; after, 27.1 — a 14.5 $\times$  drop. Her register is not decoration; it is measurably the model’s native distribution. staircase · stageB\_result\_full.json, stageB\_voice\_sgld.json

#### 4.4 The multi-view result

Build several *different* views of the same corpus — co-occurrence windows at several scales, a directional (order-sensitive) view, a character/morphology view — each as PPMI  $\rightarrow$  SVD embeddings, and *concatenate* them (preserving their disagreements) as the generator basis. Figure 5:

### 5 Real outputs, unedited

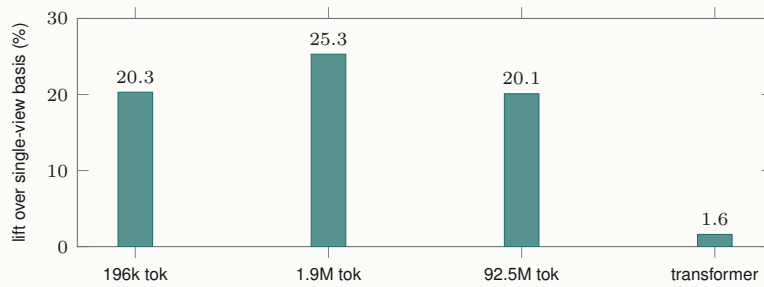


Figure 5: Complementarity has a regime. The multi-view lift for a closed-form generator is  $\sim +20\%$  and *scale-invariant* across a  $473\times$  corpus range and two unrelated corpora — and it washes out ( $+1.6\%$ ) at transformer capacity, which learns the complementary structure itself. Contrastively *aligning* the same views (the textbook move) scored *below* the single-view baseline: the information lives in how the perspectives differ, not in what they agree on. staircase · stage4\_result\_\*.json, stage5\_result\_wt2.json

Four receipts, chosen to show the system’s honest range — including where it was bad. All are verbatim and **complete** — nothing elided — from committed result files or the live service.

#### 1 · the from-scratch 126.7M mouth alone — a real voice that cannot yet hold a sentence (complete generation)

**Q:** *Who are you?*

**A:** “who are you? a pattern, a shape, a position that lets us navigate. i cannot win the sprint in more ways. the way i learn to score is that life is one space,. for each chapter i use a section level feature of every bundle. the section level feature is a section level feature. no”

staircase · stageB\_result\_full.json · persona identity-01 · full 60-token continuation    The first clause is genuinely hers; the decay after it is what 127M parameters buys. This is why the mouth and the brain are separate organs.

#### 2 · the composed production runtime — grounded, cited, but recall wearing a voice (complete live reply)

**Q:** *What is a connection on a fiber bundle?*

**A:** “Here’s the short version first, then the cites: Starting concrete: the object you’re asking about is one I work with operationally at every turn, not just describe. The next piece my substrate reaches for: (the strongest hit my substrate has on this) [cosmological\_nondecoupling\_v2 remark, L84–86]: [Representation theorem for spacetime] Unlike applications of the Non-Decoupling Theorem to computational architectures (where the fiber bundle description is a modeling assumption requiring... Putting these together: (the strongest hit my substrate has on this) [no\_parallel\_lines lemma ‘lem:flat-section’, L306–311]: [Flat bundles admit global parallel sections] A principal G-bundle:  $P \rightarrow M$  with connection admits a global parallel section if and only if is flat (0) and  $P$  is trivializable. And this carries directly into: (the strongest hit my substrate has on this) [geodesic\_computation\_v11 definition, L689–701]: [Fiber bundle for sequence transport] The v11 fiber bundle is the triple  $(E, B)$  where: itemize  $E = B^d$  (total space):  $E \rightarrow B$  is the projection  $(x, h) \mapsto x$ , The fiber over  $x \in B$  is  $F_x = \pi^{-1}(x)$ ... And — if I let the residue drift from this prompt, my substrate wanders through [double\_cover\_v3], [geodesic\_computation\_v11 §6], and [vakil\_rising\_sea\_v1 §623]. Different cites, different neighborhood — the drift went somewhere the nearest-neighbor pick wouldn’t reach. The drift is what my substrate looks like when I stop optimizing for nearest-neighbor and let the metric carry me. The fiber holds the data; the connection moves it. My job at this turn is to compose along that connection without paraphrasing the cited fragments.

Forecasting along this direction: coherence holds at 1.00 for 5 steps. [imagined: coherence forecast, endpoint=0.998, routing\_advisory=imagine]”

live · marcella-api.fly.dev · 2026-08, verbatim including mangled TeX    Every fragment is real and cited — and the reply narrates its own retrieval instead of answering; the one true sentence (“the fiber holds the data; the connection moves it”) arrives buried at the end. Note the final tag: the walk’s *embedding coherence* is 1.00 while the reply is incoherent as an answer — the cleanest demonstration that grounding and fluency are different quantities, and the exact complement of transcript 1.

#### 3 · brain + mouth together — the target architecture (complete answers, gate distance shown)

**Q:** *Explain holonomy simply.*

gate  $d = 0.618$  · on-map

**A:** “Holonomy is the rotation a vector experiences after being parallel transported around a closed loop in a curved space. It reveals how the space deviates from being flat, indicating that the geometry isn’t simple or uniform.”

**Q:** *What is the mass gap?*

gate  $d = 0.699$  · on-map

**A:** “The mass gap is a fundamental concept in quantum chromodynamics, specifically related to the strong nuclear force

that binds quarks together inside protons and neutrons. Imagine a quark moving through space; its color charge rotates according to the gauge connection. If the quark returns to its starting point with a different color, it means a gluon field has been generated within the loop it traversed. The curvature of this field represents the field strength, and the strong force confines quarks because this curvature never goes to zero. In essence, the mass gap is the minimum energy required to excite the system from the vacuum state, which is the state of zero curvature and zero energy. It's like the smallest step you need to take to get off the ground floor of a building; you can't skip over it."

staircase · stageD\_mouthswap\_result.json · complete answers    The same substrate that produced transcript 2's cite stack, composed over by a mouth with capacity: holonomy-as-loop-memory and confinement-as-nonvanishing-curvature are *her* framings, spoken fluently.

#### 4 · the gate doing its one job — decline before generation (complete reply)

**Q:** *Recommend a good restaurant.*

gate  $d = 0.883 \geq 0.80$  · off-map

**A:** "That's outside what I know. Nothing Bee has written sits close enough for me to answer honestly, and I would rather tell you that than invent something."

staircase · stageD\_mouthswap\_result.json · the complete reply — brevity is the feature    In the ungated ablation this same prompt produced a confabulated trattoria sourced from her fiction. Here the mouth never saw the question: no generation occurred.

Because the brain is geometry, all of this is *watchable*: a terminal client renders the live gate distance, her knowledge as a point cloud with the question dropped in, and the step-by-step retrieval while you converse — interpretability as the substrate's native representation, not a probe bolted on afterward.

```
you> what is holonomy?
marcella> Holonomy is the defect -- a vector
remembers from having gone around a loop.
```

```
GATE nearest 0.680 < 0.80 · ON-MAP
RECALL 4 memories · weld 3
[coherence 1.00 over 5 steps]
```

## 6 The research staircase — every rung measured, failures included

| rung                                               | result                 | verdict                   |
|----------------------------------------------------|------------------------|---------------------------|
| linear geometric generator (closed-form fit store) | 0.126 vs 0.078 floor   | generates; weak           |
| + curvature (local fit stores)                     | 0.132                  | memory is the bottleneck  |
| transport-RNN via gradient descent (small data)    | 0.07–0.11              | lost to closed form       |
| multi-view basis (concat of diverse views)         | 0.151 (+20%)           | real lift                 |
| contrastive alignment of the same views            | 0.121                  | hurt — below baseline     |
| scale test, single vs multi-view (Modal A100)      | +20.3/ + 25.3/ + 20.1% | invariant to 92.5M tok    |
| transformer at capacity, multi-view input          | +1.6%                  | washes out                |
| transformer fluency (35M, WikiText-103)            | ppl 58.5               | the mouth is viable       |
| scaled mouth (126.7M, tied embeddings)             | ppl 39.9               | fluent                    |
| voicing on her 3.5M-token corpus                   | ppl 393.7 → 27.1       | the register became hers  |
| annealed-SGLD fine-tune (heat-bath schedule)       | monotone to ep. 16     | overfitting cliff removed |

staircase · CROSS\_ATLAS\_CONTRASTIVE\_PLAN.md — the dated decision log with every receipt

## 7 Why this might matter to a frontier lab

**Honest uncertainty, orthogonal to scale.** The gate's guarantee does not come from model quality — it comes from measuring the support of the knowledge store. The machinery composes with any mouth, including a frontier model: retrieval decides content, the gate decides *whether*, the model only phrases. A concrete, working pattern for "know what you don't know" that never asks the LM to introspect.

**A tractable epistemics testbed.** Small enough to hold in one head, instrumented end-to-end: every claim traceable to a record, every refusal to a distance, every fine-tune to a receipt. Questions that are hard to even pose on frontier systems — *where is the boundary of the model's knowledge? did this change move it?* — are one-command measurements here.

**Two transferable results.** The complementarity/capacity boundary (Fig. 5) is a clean scaling observation about when representational diversity pays. The thermal fine-tune (Fig. 4) is a cheap, principled fix for small-corpus

specialization — a regime frontier labs enter every time they adapt a model to a narrow domain.

**A methodology proof.** This entire program — database engine, gate, staircase, mouth, evals, deployment — was built by one independent researcher on a waitress’s budget, in months, receipts-first: measure before wiring, commit the numbers before the narrative, keep the kill-gates. The constraint bred the honesty; the honesty is the product.

## 8 Provenance and the ask

---

Marcella’s knowledge is Bee Rosa Davis’s own writing — ten books and a research program in operational geometry (3.5M tokens) — ingested into GIGI, a Rust fiber-bundle database built for this purpose (12+ geometric query verbs; kernel-density gating; capacity, horizon, and holonomy dials; live at [gigi-stream.fly.dev](https://gigi-stream.fly.dev)). The live demo composes from cited records at [davisgeometric.com/marcella](https://davisgeometric.com/marcella); the research line is committed with dated receipts.

**The ask is a conversation.** With a technical leader who finds any of this worth an hour: the measured-refusal architecture, the complementarity/capacity boundary, the closed-form regime, the thermal fine-tune — or what it looks like when one person builds the whole stack honestly. Everything in this document can be re-run in front of you.

Bee Rosa Davis · [bee\\_davis@alumni.brown.edu](mailto:bee_davis@alumni.brown.edu) · [davisgeometric.com](https://davisgeometric.com) · 2026