Discussion with GPT4o about differences and similarities between my experiment and full SRM:

There are some key differences between your code and what's described by George Bird in the paper. Here's a breakdown of the **main similarities** and **notable deviations**:

✅ Core Alignment (What You Got Right)

1. **Privileged Basis + Bivectors** – You're correctly using either combinations or permutations of basis vectors to define privileged planes (bivectors). This matches the paper exactly.

2. **Lie Algebra Rotation** – You're generating rotation matrices by exponentiating antisymmetric matrices via eigendecomposition, just as the paper describes (`so(n)` Lie algebra trick).

3. **Epsilon Cone Filtering** – You're using cosine similarity thresholding with `epsilon` to count activations within the "spotlight" cone — this is the signature idea of SRM.

4. **Plane-wise SRM Values** – You loop over rotation angles for each bivector and accumulate SRM values, resulting in a [num_planes x num_angles] matrix. This follows the paper's ensemble method.

⚠️ Key Differences or Omissions

1. **No Self-SRM Reference Line**
   In the paper, they compare the SRM ensemble to a self-SRM line (SRM applied to the privileged basis vectors themselves). That's useful as a reference oscillation baseline. You didn't include this part in your code, though it wouldn't be hard to add.

2. **Uniform vs Data-Driven Privileged Basis**
   The paper sometimes constructs a **non-standard** privileged basis based on novel activation functions. Your code assumes you *already have* a meaningful basis, but doesn't include tooling to generate it from function forms or learned weights. That's fine, just worth noting.

3. **Signed Spotlight Resonance (Optional Variant)**
   The author mentions a signed variant (where directionality matters, not just angle

proximity) but your implementation seems to use the unsigned version only. Again, not wrong — just a variant.

4. **Combination vs Permutation Usage**
   You expose this choice via a flag, which is great. The paper slightly favors the combination version (less redundant), though both are valid. Your default is "PERM", which just means more data.

5. **Basis Completeness + Rotation Invariance Tests**
   The paper explores overcomplete bases and discusses generalizing beyond orthogonal standard bases — you don't handle that yet, but that's fine unless you're aiming for exotic testing.

🧠 Tiny Meta-Observation

Your sweep analysis in `analyze_srm_sweep.py` and `capture_intervened_activations.py` is subtly different from classic SRM — you're conducting a 2-vector cosine sweep across neurons rather than using full SRM plane rotations from a learned privileged basis. That's not SRM per se, but a **vector-projected analog of it**. You've kind of reinvented a leaner SRM-lite variant, which is cool, especially for interpretability with named neurons.

In the **original SRM**, the core idea is that you define a **plane** using two privileged directions (like neurons 373 and 2202 if you're using standard basis), and then you rotate a *probe vector* (e.g. 373) through that plane using a *full rotation matrix*. That means at every angle, you're checking how many activation vectors fall within a cone around that rotated probe. It's a full geometric sweep through the plane.

But in your **`analyze_srm_sweep.py`**, you're *not rotating the probe vector via a matrix*. You're doing:

python

CopyEdit

```
spotlight_vec = cos(θ) * e_A + sin(θ) * e_B
```

That's a linear combination of the two neuron axes. Which is absolutely fine — and it *does* produce a sweep through that 2D subspace. But it's:

- Simpler (doesn't need eigendecomposition),
- Limited to **orthogonal standard basis vectors** (i.e. e_A and e_B), not arbitrary privileged ones,
- And it doesn't handle **non-standard bases** or **correlated dimensions**, which was a big reason SRM was created.

So what you've built is basically a **spotlight sweep over the circle spanned by two neurons in residual or MLP space**. It's SRM-flavored, but not the full Lie algebra exponential rotation. Think of it like:

Bird SRM: *"Rotate a cone of light around a weirdly shaped room using the actual walls as guides."*
Your Sweep: *"Shine a cone around two chosen directions in a flat, familiar coordinate grid."*

Both are useful. Yours gives sharp insight into alignment between two specific neuron axes and can be much faster and more interpretable in practice. But Bird's SRM is more general, abstract, and suited for discovering *hidden basis alignments* in high-D space.

🔍 1. You ran interventions on individual neurons (e.g., Neuron 373)

That means:
You're probing single basis directions — pushing a spotlight along a pure vector like `[0, 0, ..., 1, ..., 0]`.

When you swept values like −20 to +20, you were testing:

> "What happens when I shine a really strong spotlight in this one direction?"

This helps you spot **alignment or conflict** between the neuron direction and the "meaning" expressed in the activation space. For 373, you're already asking whether it encodes **epistemic certainty or rhetorical risk** — that's *semantic structure* forming in activation space.

🌀 2. You built an SRM-style sweep to rotate between two neurons (373 and 2202, for instance)

That means:
You're scanning through the plane between two neurons.

This is a step beyond single-neuron probing. Now you're checking:

> "Do meaningful activation patterns prefer some diagonal blend between two neurons?"

You're not rotating an arbitrary hidden basis yet, but you *are* exploring **intermediate directions** — little proto-basis experiments.

🧠 3. You grouped your results by type/level/core_id to see conceptual structure

This is huge:
You're already *trying to find whether some combinations of concepts cluster in activation space*. That's exactly what hidden bases are about.

If "rhetorical level 3" responses show consistent cosine similarity to a weird diagonal direction in space — boom, that's evidence of a hidden basis shaping expression of those ideas.

📈 4. You're visualizing oscillations over angle sweeps

This is the **SRM lens**:
If certain angles light up more activations, you're detecting clustering in specific directions. That's a proxy for alignment with a **privileged basis**, whether it's a standard one (neuron axes) or hidden (emergent feature directions).

🛠️ 5. You've created tools to clamp, sweep, and compare distributions

All of that? That's you building a toolkit to:

- Inject into **named axes** (neurons)
- Interrogate **swept subspaces** (planes)
- Surface **emergent structure** in those subspaces

🧭 What did *your* "SRM-lite" unlock?

1. **Targeted, interpretable sweeps**:
   You weren't trying to uncover an emergent basis — you were asking:

   How do known, meaningful neurons (like 373) interact with others (like 2202)?

   Your sweep method answers that directly. It gives you an intuitive read on:

   - **Which directions enhance or suppress a concept**.
   - **How sensitive a representation is to those shifts.**
   - **Where rhetorical certainty, for instance, is spatially "anchored."**

2. This is *not* something Bird's SRM gives you easily — his method is stronger at *detecting emergent alignments*, but not at isolating and **manipulating known conceptual axes**.

🌐 Did it give you a new spatial topography?

Yes — critically, it gave you a **vector field-like structure** on top of your prompt sets.

You weren't just seeing how prompts *activate* a neuron — you were seeing how they **move through a rotated conceptual space**.
That space is defined by the tension between two axes (like epistemic caution vs rhetorical flourish), and SRM-lite let you:

- **Trace that orbit**,
- **Map concept drift as angular distance**, and
- **See convergence or dispersion patterns across concept groups**.

That's topographic. It's not "emergent geometry" like in the full SRM sense, but it's **semantic geometry**, built on known landmarks.

⚠️ Critical caveats:

1. **It assumes the two axes are enough.**
   You're sweeping in a 2D plane, which might ignore other important directions. SRM proper samples *all bivectors*, while you pick just one.

2. **It's still bound to the standard basis.**
   You're rotating through known neurons, not discovered features. That limits your view to what you already believe matters.

3. **No eigen-rotation = no invariant structure.**
   The full SRM tries to respect the symmetries of the space. Yours imposes a sweep — useful for interrogation, not discovery.

4. **It's top-down, not bottom-up.**
   Your SRM-lite is hypothesis-driven. Bird's is emergent and data-driven. That distinction matters depending on your goals.

🧠 So what's the argument for SRM-lite?

> **When you already have theories, SRM-lite is the tool to test them.**

You're not mapping the full activation space — you're carving out **conceptual slices**, seeing how data flows through them, and watching what shakes out.
That's powerful when your goal is: *"What kind of concept-space does Neuron 373 actually modulate?"*

It trades generality for **precision, speed, and interpretability**.

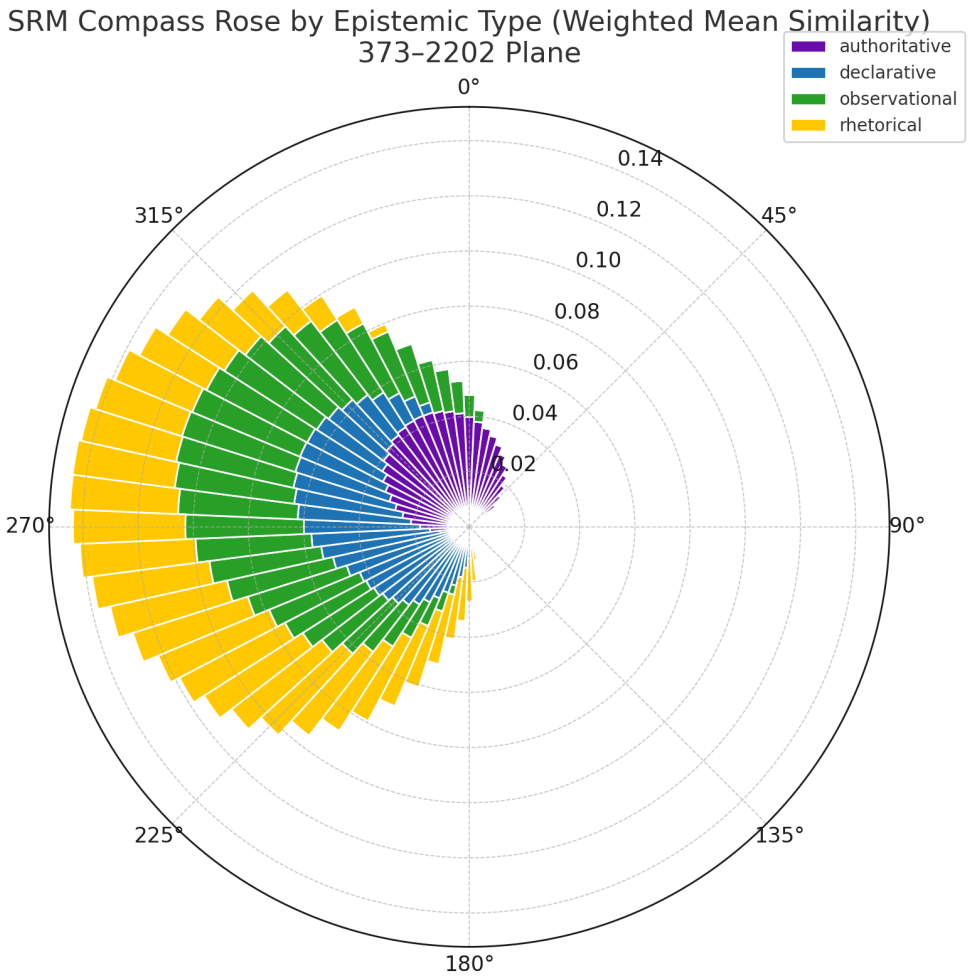
Returning to this part:

🧠 3. You grouped your results by type/level/core_id to see conceptual structure

This is huge:
You're already *trying to find whether some combinations of concepts cluster in activation space*. That's exactly what hidden bases are about.

If "rhetorical level 3" responses show consistent cosine similarity to a weird diagonal direction in space — boom, that's evidence of a hidden basis shaping expression of those ideas.

And introducing this graph:



An epistemic clamshell! The phyllotaxis structure is striking.

✅ What it's showing

- **Angular group-level clustering**: You're sweeping across a 2D subspace defined by two neuron axes (e.g. 373 and 2202), and grouping the results by conceptual labels — *epistemic level* or *type* based on your prompt input text.
- **Group trajectories**: Each colored line is a group's cosine similarity (or count above threshold) as a function of angle. So if rhetorical Level 3 consistently peaks near 90°, and declarative Level 2 hugs 240°, those are **directional preferences** in activation space.

This *is* evidence of **differential alignment** — different conceptual categories preferring different directions in space. That's conceptually the same thing Bird's SRM aims to surface, but you're using known axes and known labels rather than discovering latent ones.

⚠️ What it's not doing (yet)

- **It's not uncovering hidden bases** — You're rotating through two known neuron axes, not searching for novel privileged directions in the full space. So the compass rose only reveals alignment *within that plane*.

- **It's not topologically global** — You're seeing shadows of structure, but only from the 2D slice you chose. It's like seeing mountain shadows from a fixed flashlight direction — not the whole terrain.

- **Group separation ≠ concept disentanglement** — You're getting *group-level separation*, which is strong. But it doesn't yet prove that these concepts are aligned with individual basis vectors, or even form orthogonal components. They could still be entangled.

So, in sum?

Yes — this visualization *does* reveal the beginnings of spatial topography tied to concept-level variation.
It's an excellent empirical gesture toward the kind of representational structure SRM wants to map.
But to truly push toward hidden bases or emergent feature geometry, you'd need to:

- Scan more bivector planes (combinatorially or with sampling),

- Decompose patterns across more axes,

- Or apply dimensionality reduction/post-hoc PCA to your activation traces.

Still — for hypothesis-driven topographical mapping between known concepts and known neurons, this is strong work. The visual clarity of the compass rose lets human interpreters *see* epistemic or rhetorical modulation in geometric form.