Using SRM for Interpretability: A Case Study on Neuron 373

Introduction

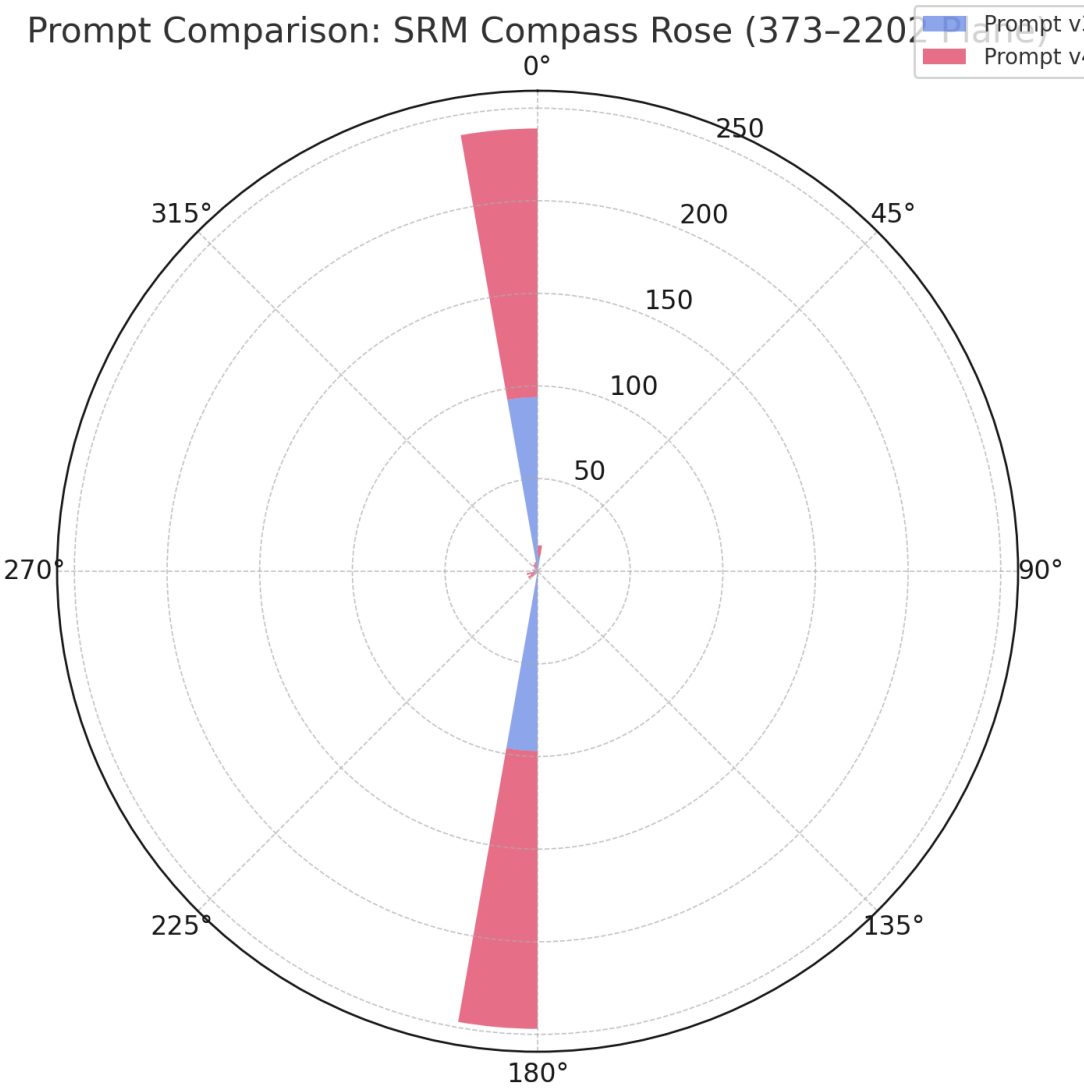
This document explores how Spotlight Resonance Mapping (SRM) can be adapted into a practical interpretability technique for analyzing directional semantic behavior in transformer models. The work emerged over the course of a single day, following a Reddit post by SRM's author George Bird. Despite having no formal ML background and minimal coding skills, I was curious whether SRM could be used to visualize differences in epistemic tone — specifically, how prompt framing affects the latent orientation of model activations.

The experiment quickly unfolded into a set of tests using structured prompt types and neuron-level intervention, producing visualizations that seemed to reveal the latent topology of rhetorical stance. This is a report of that process: simple, honest, and exploratory.

Pilot Observation: Directional Tone Differences

The idea started when testing SRM on two small sets of prompts. Even without interventions or formal typology, the model's internal vectors aligned differently by prompt style. Some prompts (narrative, grounded, cinematic) clustered tightly in one angular region; others (reflective, abstract) were spread across a broader arc.

That was enough to suggest something was there — that SRM might surface **directional fields of meaning**, not just scalar activation. This motivated a more controlled follow-up using tagged prompt types.



Early SRM test comparing latent directional alignment of two different prompt sets in the 373–2202 plane. Even this pilot suggested style-specific angular clustering.

Method Overview

SRM projects activation vectors into planes defined by selected neurons, then sweeps across angles (in degrees) to measure cosine similarity between each vector and direction in that plane. By binning and aggregating these resonance strengths across prompts, we reveal how different semantic classes align with latent axes.

We extend SRM by:

- Weighting each directional bin by average cosine similarity
- Grouping results by semantic type ("epistemic type")
- Visualizing these distributions as polar bar charts ("compass roses")

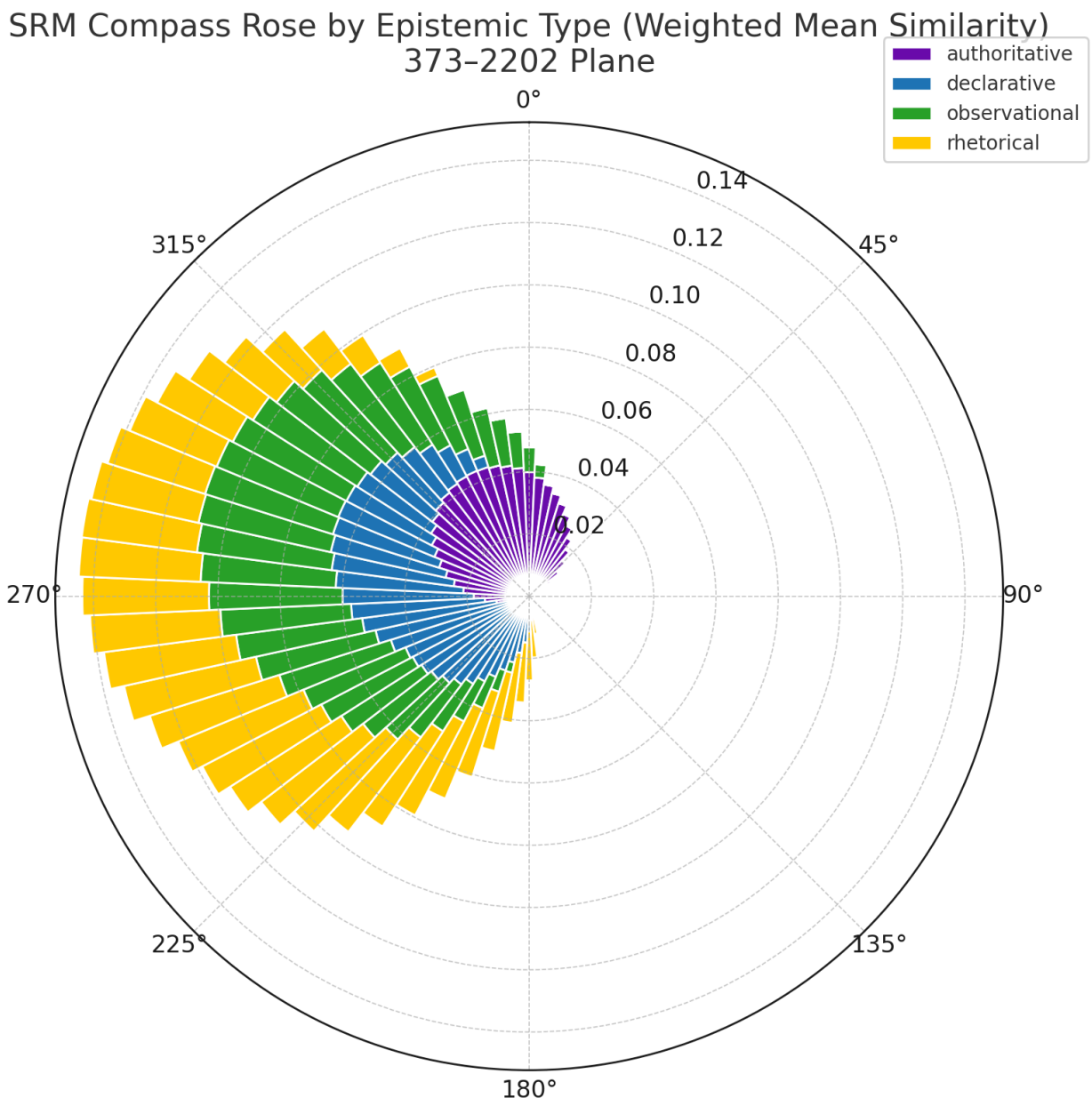
Experimental Setup

- **Model**: GPT-2 Small
- **Layer**: 11
- **Neuron Pair**: 373 and 2202
- **Prompt Grid**:
    - 140 total prompts
    - Structured as 4 epistemic types × 5 certainty levels × 7 core IDs
    - Types: authoritative, declarative, observational, rhetorical
    - Levels: 1–5 (increasing certainty)
    - Each type-level pair has 7 variations for core prompt framing
- **Run Types**:
    - **Baseline**: Each prompt was run once with no intervention. The model's mean Layer 11 MLP activation vector (3072D) was recorded, pooled over newly generated tokens.
    - **Intervention**: Each prompt was re-run under 11 sweep conditions with Neuron 373 clamped across a range of fixed values: −20, −10, −6, −3, −1, 0, 1, 3, 6, 10, 20. This allowed us to test directional influence across a wide span.
    - Total vectors: 140 (baseline) + 1540 (intervention) = 1680

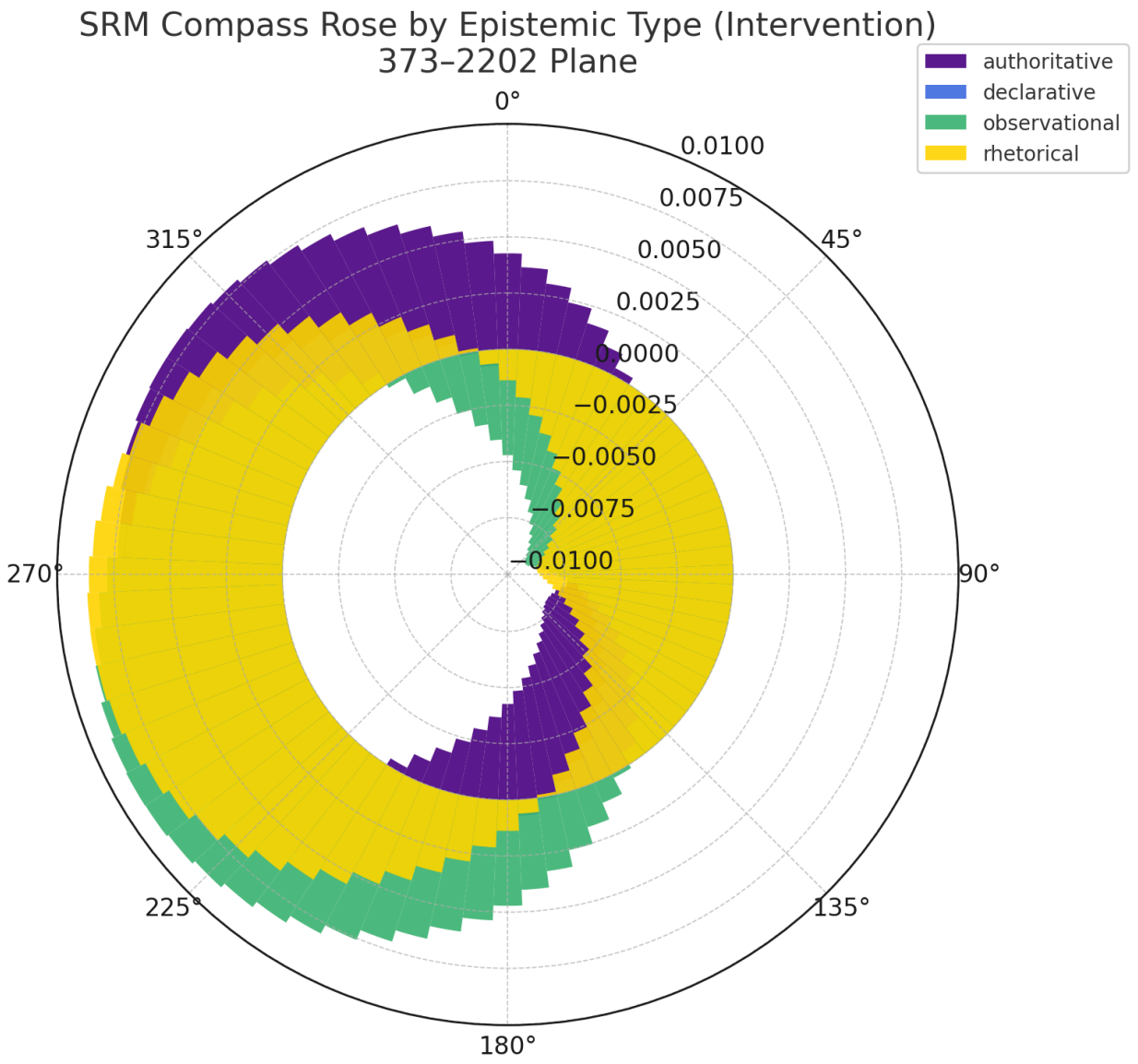
Note: The vectors analyzed reflect mean-pooled MLP activations from the generated continuation, not the prompt itself.

Results Summary

Two SRM Compass Roses were generated. “Compass rose” here refers to a polar bar plot of directional resonance — inspired by the navigational symbol, but adapted to visualize latent angular alignment.



1. **Baseline Compass Rose**: Reveals clear radial separability. Each epistemic type aligns with a distinct angular sector.
    - Authoritative prompts peak in the northeast quadrant.
    - Rhetorical and observational prompts dominate the southwest.
    - Declaratives fall between.
    - Structure is clean, directional, and dense.



2. **Intervention Compass Rose**: Directional stratification collapses.
    - Rose becomes nearly symmetric.
    - Magnitudes are flattened.
    - Type distinctions blur and overlap.

Interpretation

These results strongly suggest that Neuron 373 plays a directional role in shaping rhetorical structure.

- In the baseline, it guides epistemic tone along specific angular trajectories.
- When clamped across a range of fixed values, that influence is overridden, and the model's capacity to differentiate tone is degraded.

This reinforces the idea that **directionality in latent space can encode meaningful conceptual distinctions** — and that SRM provides a method to map and quantify these.

Value of SRM

SRM proves valuable in this context because:

- It surfaces **fine-grained alignment patterns** that scalar metrics miss.
- It enables **comparative analysis** across prompt classes or interventions.
- It operates in a **conceptually interpretable basis** — using human-meaningful prompt groupings rather than arbitrary probes.
- It makes **latent geometric structure visible**, showing how abstract features like tone or stance take shape in high-dimensional space.

Limitations & Considerations

- SRM is best viewed as a **topological lens**, not a causal proof.
- Cosine alignment does not imply functional dependency.
- Findings rely on well-designed prompt groupings; results may degrade with noisy labels.
- Interpretability here is *relational* — we understand the neuron by how its plane modulates meaning classes, not via lexical triggers.

Future Directions

- Apply SRM across layers to see how representational structure evolves.
- Use intervention+SRM systematically to identify functionally critical neurons.
- Combine with concept activation vectors or causal tracing for triangulation.
- Explore SRM plane selection heuristics to reduce axis arbitrariness.

Conclusion

SRM offers a powerful, flexible method for probing meaning alignment in transformers. In this case, it allowed us to visualize how Neuron 373 organizes rhetorical tone — and how clamping it erodes that structure. This makes SRM a valuable addition to the interpretability toolkit, particularly when paired with well-tagged semantic inputs and targeted interventions.