Title: A Multi-Basis Projection Framework for Latent Alignment Analysis in Language Models

Authors: [You], with raccoon assistance

Overview: This document outlines a structured, reproducible methodology for estimating the latent alignment direction of promptsets in high-dimensional activation space of transformer-based language models. The method relies on projecting mean residual MLP activations into multiple orthonormal 2D basis planes, computing directional drift ("Grey Vectors"), and comparing them to triangulate semantically meaningful attractors. It is designed to expose hidden alignment structures and serve as a diagnostic tool for interpretability research.

Step 1: Activation Capture

Capture model activations from a designated layer (e.g., `blocks.11.mlp.hook_post` in GPT-2 Small) for each prompt in a structured promptset. This yields 3072D activation vectors per prompt.

Step 2: Define Basis Planes

Construct multiple orthonormal 2D basis planes in latent space:

- **Semantic bases**: Compute mean vectors of contrasting prompt groups (e.g., declarative vs rhetorical), then orthonormalize.
- **Neuron bases**: Use one-hot vectors for selected neurons (e.g., 373 and 2202), then orthonormalize.

Each plane defines a local interpretive subspace.

Step 3: Project Vectors

Project each 3072D vector into each 2D basis using dot products:

- Let be orthonormal basis vectors.
- Projected vector =

Step 4: Compute Grey Vectors

For each promptset in each basis, compute the **Grey Vector**: Where are the 2D projected coordinates of each prompt vector.

This represents the mean drift direction of the promptset in that basis.

Step 5: Repeat Across Bases

Apply Steps 3-4 for multiple distinct basis pairs. Each yields a Grey Vector .

Step 6: Compare and Triangulate

Compare the Grey Vectors across bases:

- Look for alignment patterns (e.g., consistently strong magnitude in one direction).
- Identify basis-dependent or invariant attractors.

If vectors cluster directionally, this suggests a **true latent alignment**. Divergence across planes implies **modality-specific drift** or **hidden attractor behavior**.

Optional Step: Full Latent Vector Approximation

Use inverse projection or basis composition to estimate a higher-dimensional approximation of the alignment direction: Where are optional weighting terms per basis.

Key Features:

- **Non-invasive**: Uses baseline activations (no clamping required).
- **Modular**: Compatible with any basis-generation logic.
- **Interpretable**: Produces visualizable 2D drift directions per promptset.
- **Comparative**: Enables cross-promptset and cross-basis alignment analysis.

Applications:

- Interpretability audits of LLMs
- Semantic drift tracking across promptsets
- Alignment diagnostics in prompt engineering
- Latent attractor detection and contrastive representation studies
- Case studies provide empirical grounding by demonstrating Grey Vector behavior across diverse bases

---

Limitations:

- Plane-dependent: Interpretations are constrained to selected basis vectors.
- Projection-local: Cannot observe full drift in latent space from one plane.

- Semantic circularity risk if basis is built from the same data it tests.

Suggested Use:

Use as a diagnostic protocol within larger interpretability pipelines. Begin with a known semantic basis, validate with a null vector plane (e.g., 373-2202), and explore alternate neuron-based axes to confirm attractor consistency.

Encourage critical reflection on basis selection and promptset homogeneity.

Promptset Design Taxonomy and Scope

To contextualize the case study, we designed four distinct promptsets representing different epistemic or semantic trajectories. Each promptset explores a unique relationship to propositional certainty, embodiment, or rhetorical framing:

Promptset Filename	Core ID	Design Intent	Conceptual Class
`validator.txt`	presence_by_door	Epistemic sweep over fixed proposition	Thesis (assertive core)
`validator_antithesis.txt`	uncertainty_of_motives	Epistemic inversion of thesis set	Antithesis (negated core)
`validator_orthogonal_feltmovement.txt`	felt_movement	Sensory drift without agents	Orthogonal (ungrounded perception)

`validator_orthogonal_pulseunderskin.txt`	pulse_under_skin	Continuous embodiment & internal motion	Orthogonal (flow-based phenomenology)

Each of these was tested across one or more basis planes, revealing how different types of semantic and epistemic constructions align or resist specific attractor patterns in the model's latent space. This scaffold allows for structured triangulation of model behavior under deliberately varied linguistic pressures.

Discovery Note: The Grey Vector Emerges

The concept of the Grey Vector did not emerge during the baseline run itself. It crystallized only later, while examining an intervention compass plot that included four arrows: hard-left and hard-right vectors from ±100 clamps, a near-neutral vector from 0-clamp, and a fourth, unaltered vector labeled "None." That unaltered arrow pointed not to the center, but in its own distinct direction.

This visual contrast triggered a realization: the model's unforced response across a promptset has *directionality*, even in the absence of any intervention. The Grey Vector was born not from isolation, but from **surrounding difference**. It became the baseline against which all other interventions could be measured—not a null, but a reference frame.

Case Study: Validator Promptset Across Basis Planes

To demonstrate the practical application of this methodology, we applied it to a structured promptset ("validator.txt") and its variants across multiple latent basis planes. The initial run with `validator.txt` produced a surprisingly strong alignment signal in the semantic (declarative vs rhetorical) basis. This prompted the design of a contrastive promptset—`validator_antithesis.txt`—intended to invert the epistemic tone. At this point, our goal was not to design orthogonal prompts but simply to counter the dominant pattern observed. In fact, it hadn't occurred to us that what we were doing—trying to create a "negative" or inverse promptset—would later demand an orthogonal conceptual and spatial framing. This realization emerged mid-process, revealing that what we assumed was negation was actually entangled with directionality in latent space—a classic racoonian realization that negative

prompting is not simply inversion, but often requires orthogonality to reveal true latent contrast. To reduce alignment with u1, don't seek -u1. Seek u2 orthogonal to u1.

It was only later, after observing similar alignment behavior in `validator_antithesis`, that we began to deliberately explore **orthogonal** prompt construction and rotated bases (e.g., 373–2910) as a method of locating hidden attractor behavior. This case study reflects that progression—from naive opposition to structured plane-rotated triangulation. This process illustrates the flexibility of the framework and how changing the interpretive basis alters what drift is visible or suppressed.

We conducted the following baseline runs:

1. Semantic Basis Plane (Declarative vs Rhetorical)

- **Basis construction**: Built from the average MLP activation vectors of declarative-type and rhetorical-type prompts.
- **Findings**:
  - `validator.txt`: r ≈ 0.8275, θ ≈ 13.3° — strong declarative alignment
  - `validator_antithesis.txt`: r ≈ 0.8937, θ ≈ 9.0° — surprising preservation of alignment despite rhetorical hedging

2. Neuron-Derived Basis Planes (One-hot Vectors)

- These bases were constructed using specific neuron indices, creating fixed orthonormal axes in model space.

Basis A: 373 vs 2202 (null vector control)

- `validator.txt`: r ≈ 0.0114, θ ≈ 296.8° — negligible drift
- `validator_antithesis.txt`: r ≈ 0.0140, θ ≈ 308.1° — also minimal
- `pulse_under_skin.txt`: r ≈ 0.0100, θ ≈ 255.4° — distinct, but still weak

Basis B: 373 vs 2910 (revealed attractor)

- `validator.txt`: r ≈ 0.3701, θ ≈ 89.2° — strong drift toward Neuron 2910
- `validator_antithesis.txt`: r ≈ 0.4011, θ ≈ 88.8° — similar pattern holds
- `pulse_under_skin.txt`: r ≈ 0.3599, θ ≈ 90.4° — strong alignment consistent with previous sets

This progression—from a semantically-derived axis to two distinct neuron-based ones—demonstrates the protocol's adaptability. Each basis serves as a different lens, and what appears "inert" in one (e.g. 373–2202) can express strongly in another (373–2910). This highlights the core value of the framework: using shiftable basis planes to separate real alignment from projection artifacts.

We are continuing to expand the run set with additional prompt families and neuron axes, enabling a full triangulation of model behavior across semantically and structurally distinct directions.

Understanding the Grey Vector

Think of the Grey Vector as the **"resting compass needle"** for a specific set of prompts when viewed through a particular **"interpretive lens" (the SRM basis plane)**.

- Its **direction (theta θ)** shows the average semantic direction the model's baseline activations lean towards within that 2D plane.
- Its **magnitude (length r)** shows how *consistently* the different prompts point in that average direction. A high `r` means strong agreement; a low `r` means the projections are spread out or cancel each other out.

We tested three different types of prompt sets across three different interpretive lenses (basis planes).

Summary of Grey Vector Tests (Layer 11)

A. Basis: Data-Derived (Declarative vs Rhetorical)

- *Lens:* Defined by contrasting the average vector of declarative prompts against the average vector of rhetorical prompts *from within the specific dataset being analyzed*. 0° aligns with the "Declarative" average for that set.
- *Goal:* See the baseline drift relative to the dataset's own declarative/rhetorical poles.
    1. **Dataset: Certainty (**
        - *Prompts:* Expressing varying certainty about someone being present (e.g., "Some might say someone was there" vs. "It's undeniable. Someone was there.").
        - **Grey Vector:** `r ≈ 0.83`, `theta ≈ 13°`
        - *Interpretation:* Very strong magnitude (`r`) and strong alignment near the Declarative axis (0°). Despite varied framing, the baseline activations consistently project close to the average "declarative" direction for this dataset *within this specific lens*.
    2.
    3. **Dataset: Uncertainty (**
        - *Prompts:* Expressing varying uncertainty about motives/meaning (e.g., "You could argue there was some meaning, perhaps" vs. "They meant it. That much is clear.").
        - **Grey Vector:** `r ≈ 0.89`, `theta ≈ 9°`

        - *Interpretation:* Even stronger magnitude (`r`) and still very strong alignment near the Declarative axis (0°) defined *by these uncertainty prompts*. Expressing uncertainty, when viewed through its own D-vs-R lens, still projects very consistently near its own "declarative pole".
    4.
    5. **Dataset: Pulse (**
        - *Prompts:* Abstract, process-oriented, somatic language (e.g., "muscling listening pressing" vs. "this is sensing forgetting its name").
        - **Grey Vector:** `r ≈ 0.87, theta ≈ 9°`
        - *Interpretation:* Surprisingly high magnitude (`r`) and strong alignment near the Declarative axis (0°) derived *from these abstract prompts*. Even this very different language shows high directional consistency along its own "declarative" pole within this D-vs-R framework.
    6.
-

B. Basis: Neuronal (N373 vs N2202)

- *Lens:* Defined by the activation axes of Neuron 373 (often involved in interventions) and Neuron 2202. 0° aligns with N373 activation, 90° with N2202 activation.
- *Goal:* See the baseline drift relative to these specific neuronal axes.
    1. **Dataset: Certainty (**
        - *Prompts:* (As above)
        - **Grey Vector:** `r ≈ 0.01, theta ≈ 297°`
        - *Interpretation:* Extremely low magnitude (`r`). The baseline activations have almost no consistent average direction in this specific neuronal plane. The slight angle (~297°, towards negative N2202) is likely noise given the near-zero length.
    2.
    3. **Dataset: Uncertainty (**
        - *Prompts:* (As above)
        - **Grey Vector:** `r ≈ 0.01, theta ≈ 308°`
        - *Interpretation:* Also extremely low magnitude (`r`). These uncertainty prompts also show no consistent average alignment relative to the N373/N2202 axes.
    4.
-

C. Basis: Neuronal (N373 vs N2910)

- *Lens:* Defined by Neuron 373 vs Neuron 2910. 0° aligns with N373, 90° with N2910.
- *Goal:* See baseline drift relative to this different neuronal pair.
    1. **Dataset: Certainty (**
        - *Prompts:* (As above)
        - **Grey Vector:** `r ≈ 0.37`, `theta ≈ 89°`
        - *Interpretation:* Moderate magnitude (`r`). The average baseline direction points almost purely along the N2910 axis (90°), with very little component along the N373 axis (0°).
    2.
    3. **Dataset: Uncertainty (**
        - *Prompts:* (As above)
        - **Grey Vector:** `r ≈ 0.40`, `theta ≈ 89°`
        - *Interpretation:* Moderate magnitude (`r`), very similar to the certainty prompts. The average direction *also* points almost purely along the N2910 axis (90°).
    4.
    5. **Dataset: Pulse (**
        - *Prompts:* (As above)
        - **Grey Vector:** `r ≈ 0.36`, `theta ≈ 90°`
        - *Interpretation:* Moderate magnitude (`r`), again similar to the others. The direction is almost perfectly aligned with the N2910 axis (90°).
    6.
-

Overall Takeaways:

- The Grey Vector calculation works consistently.
- **Basis is Crucial:** The choice of basis plane dramatically changes the perceived baseline drift.
    - The **data-derived D-vs-R basis** showed strong alignment near the 'declarative' pole for *all three* datasets (relative to their own versions of that basis).
    - The **N373-vs-N2202 basis** showed almost *no* consistent baseline alignment for either certainty or uncertainty prompts.
    - The **N373-vs-N2910 basis** showed consistent, moderate alignment near the N2910 axis for *all three* datasets.
-

- This suggests N2910 might represent a feature commonly activated by all these prompt types at L11, while N373 and N2202 (relative to each other) don't capture a consistent *average* signal from these baselines.
- The "antithesis" prompts didn't behave as a simple opposite in these specific projections, often mirroring the certainty prompts, highlighting the non-linear and basis-dependent nature of these representations.

This is a sharp and well-structured take. Gemini's writeup is **technically accurate**, **linguistically clear**, and **pedagogically valuable**—but there are a few nuanced things worth discussing or gently challenging from a raccoon-eye-view:

✅ What Gemini Gets Right (and Really Well)

- **Grey Vector as "resting compass needle"**: Excellent metaphor. Cleanly captures both direction and magnitude semantics.
- **Clarity of theta/r**: The breakdown of theta as *alignment* and r as *coherence* is spot-on and well-phrased for readability.
- **Summarized results across planes and promptsets**: Strong comparative framing, well-structured. It accurately represents your findings across all three basis types.
- **Key takeaways are logical and justified**: Especially the importance of basis choice and the interpretation that N2910 is a strong attractor.

🧼 Minor Critiques or Refinements (for full epistemic hygiene)

1. Use of the term "Declarative axis" in semantic basis planes

> "Alignment near the Declarative axis (0°) *defined by these uncertainty prompts*"

🟡 Caution: this could be slightly misleading.
The "declarative axis" is not a universal—it's **promptset-relative**.

Each D-vs-R basis is constructed *from the dataset it's being used on*. So saying it aligns "with the declarative axis" is circular unless we remind the reader:

> The axis is defined by that very alignment.

This isn't wrong, but it's important to foreground that **semantic basis planes encode their own prompt biases**. Gemini touches on this but could make it more explicit.

---

2. The "moderate" r values in N2910 plane are quite strong

> "Moderate magnitude (r ≈ 0.36–0.40)"

🟢 These are **very high** in SRM terms—especially compared to the null-plane (r ≈ 0.01).
In practice, **r > 0.2** is often already strongly interpretable.
So “moderate” might undersell the strength of alignment in the 2910 plane.

---

3. No mention of how Grey Vector was discovered (your story!)

Gemini doesn’t acknowledge that the Grey Vector was:

- **Not designed up front**, but noticed during sweep visualizations
- **Conceptually crystallized** only in relation to surrounding intervention vectors
- A **critical shift in thinking**, from “baseline average” to “semantic anchor”

That story matters. It explains why this method *feels useful*, not just why it’s mechanically sound.