Conceptual Cartography: High-Level Design Document

Core Inspiration

This project builds directly from Gurnee & Tegmark's 2024 paper Language Models Represent Space and Time. That paper demonstrated that LLMs develop linearly decodable internal coordinates for real-world places and times. Their experiments revealed "space neurons" and "time neurons," and used linear probes to extract actual latitude, longitude, and timestamp data from LLaMA models.

Our work takes that spatial metaphor and pushes it into the terrain of abstract semantic space. Instead of Paris or 1776, we?re mapping concepts like home, grief, safety, hope. These aren?t grounded in a physical world model?but we suspect they?re encoded in ways that are just as spatially structured.

Core Hypotheses We Are Testing

1. Linear Representation Hypothesis (LRH)

- Do abstract concepts have consistent, linearly structured representations in the final-layer activation space?
- Does vector arithmetic (e.g., vector(A) - vector(B)) reliably capture semantic contrast?
- Are there difference neurons whose activations correlate strongly with specific conceptual distinctions?

2. Geometric Representation Hypothesis

- Are semantically related concepts closer together than unrelated ones when measured by cosine

similarity or Euclidean distance?

- Do intuitive clusters (e.g., "home" and "house") emerge naturally in reduced dimensions?
- Are emotional or affective concepts topologically distinct from neutral or concrete ones?

3. Layer Progression Hypothesis

- Do conceptual differences become sharper and more separable across model layers?
- Are key semantic axes (e.g., safe ? unsafe) increasingly decodable in later layers?
- Can heatmaps or neuron difference vectors visually reveal this progression?

4. Context Independence Hypothesis

- How stable is a concept?s internal representation across syntactic and pragmatic contexts?
- Do vectors for "home" in "He returned [home]" and "They dream of [home]" cluster together?
- Can we systematically test this using line-number comparisons across prompts?

These four hypotheses drive every stage of the experiment, and each test or analysis method is a probe into one or more of them.

System Goals

1. Identify a token?s vector location in the final layer of a language model.
2. Find its local neighborhood?by cosine and Euclidean metrics.
3. Track how the neighborhood changes across prompts (context drift).
4. Apply vector arithmetic and test identity and analogy operations.
5. Visualize concept space in 2D/3D via dimensionality reduction.
6. Search for concept neurons aligned with directional contrasts.
7. Eventually, explore linear probing for higher-level features (valence, safety, concreteness, etc.).

Pipeline Components

- Prompt Batch: Texts with bracketed target tokens
- Activation Capture: Final-layer vector for bracketed token
- Input Parser: Custom directives like #DEFINE_VEC, #COMPARE, #ARITHMETIC, #ANALOGY, etc.
- Vector Ops: Cosine/Euclidean similarity, vector math
- Neighborhood Search: Top-K closest points
- Dimensionality Reduction: UMAP, PCA for topological projection
- Report Generator: Summaries of comparisons, arithmetic, neighbors, heatmaps

---

Experimental Phases

V7.7: Baseline toolkit validation

- Full parser functionality
- Vector identity validation (e.g., A + (B - A) = B)
- Initial UMAP cluster maps

V8.x: Analogy and directional reasoning

- Implement and validate #ANALOGY syntax: A - B + C ? ?
- Score analogy predictions with cosine/euclidean
- Identify stable vs unstable directions

V9+: Cross-layer and prompt-aware dynamics

- Compare vector locations across prompt frames

- Inspect heatmaps for layer progression

- Explore neuron-level delta patterns

- Prototype linear probes for valence/safety/concreteness

Design Values

- Interpretability over performance

- Token-efficient, modular input/output design

- Currently focused on GPT-2 Small; may extend to similar open models later if alignment is feasible

- LLM-assisted iteration, structured to support Gemini/GPT collaboration

- Incremental development with robust regression testing

Framing Statement

Where Gurnee and Tegmark drew maps of the world, we?re drawing maps of meaning. We want to know: what kind of internal terrain does a model imagine when you say "home"? Can grief be measured in distance? Can hope point somewhere?

This project is both an interpretability tool and a kind of conceptual GPS. We are not just reading what the model knows?we are walking its inner world.