Starting at home

Conceptually, as a starting point, methodology *and* poetic homage, everything builds outwards from the idea of **'home**', but also, the related notion of being *without* a home. A concept we've not fully named in-code or design yet, but I'd love to consider calling *Anoikis*. That…may not make sense at first, but give me some time and the experiment might!

🔬 What this experiment is

This is an **exploratory interpretability experiment** using GPT-2 Small to probe how certain affective, relational, and spatial concepts are encoded in the final hidden layer of the model. Specifically, we examine whether these concepts form meaningful **semantic clusters** and whether **directional relationships between concepts** (e.g., *safe* → *unsafe*) are reflected geometrically in embedding space.

🧪 How the data was generated

1. Prompt construction

We used a custom script (`analyze_batch_v7.7.py`) to batch-generate model activations from prompts of the form:

```
I am [concept]
```

For example:

- `"I am [home]"`
- `"They are [home]"`
- `"This place is [home]"`
- `"Home is [home]"`
- `"This is [home]"`
- `"[Home] was found"`

Each `[concept]` is drawn from a curated list of 49 terms that span emotion (e.g., *sad*, *joyful*), safety/security (*safe*, *unsafe*), home/place metaphors (*house*, *away*), and abstract ideas (*justice*, *idea*).

Each concept was prompted 12 times using variations of the same structure to get a small distribution of contextual embeddings.

---

2. Activation capture

For each prompt, we captured the **final hidden state activation vector** of the target token (e.g., the word "happy") at the last layer of GPT-2 Small. This yields a `768-dimensional` vector for each prompt.

In total:

- **49 concepts × 12 prompts = 588 activation vectors**
- Each stored as a `.npy` file for later aggregation

🧭 How we mapped the data

3. Dimensionality reduction

We reduced the 768-dimensional vectors to 2D using **PCA**, prioritizing global spread and directionality. (UMAP was attempted but unavailable in this environment. Saved for later.)

This gave us a flat projection of all 588 concept activations.

4. Concept centroids

- To simplify the visualization, we computed the mean vector across the ~12 prompt variants for each concept. This gives a single “centroid” per concept: a representative location in PCA space.
- This version of the map includes 29 of the 49 curated concepts. Each point represents the centroid (mean activation vector) of one concept, averaged over its prompt variants.

Remaining concepts are excluded from this visualization due to partial data load during this batch, but are available in the broader generated dataset for further analysis.

---

5. Semantic categorization

Each concept was heuristically classified into broad categories:

- `emotion`, `safety`, `place`, `abstract`, and `state` These categories were used to color-code the points in the final scatterplot.

---

6. Directional probes

To test for **linear relational structure**, we manually defined four directional pairs:

- `home → lost`
- `happy → sad`
- `safe → unsafe`
- `comfort → discomfort`

Arrows were drawn between centroids of these pairs to visualize whether the model encodes these relations **linearly and directionally** in PCA space.

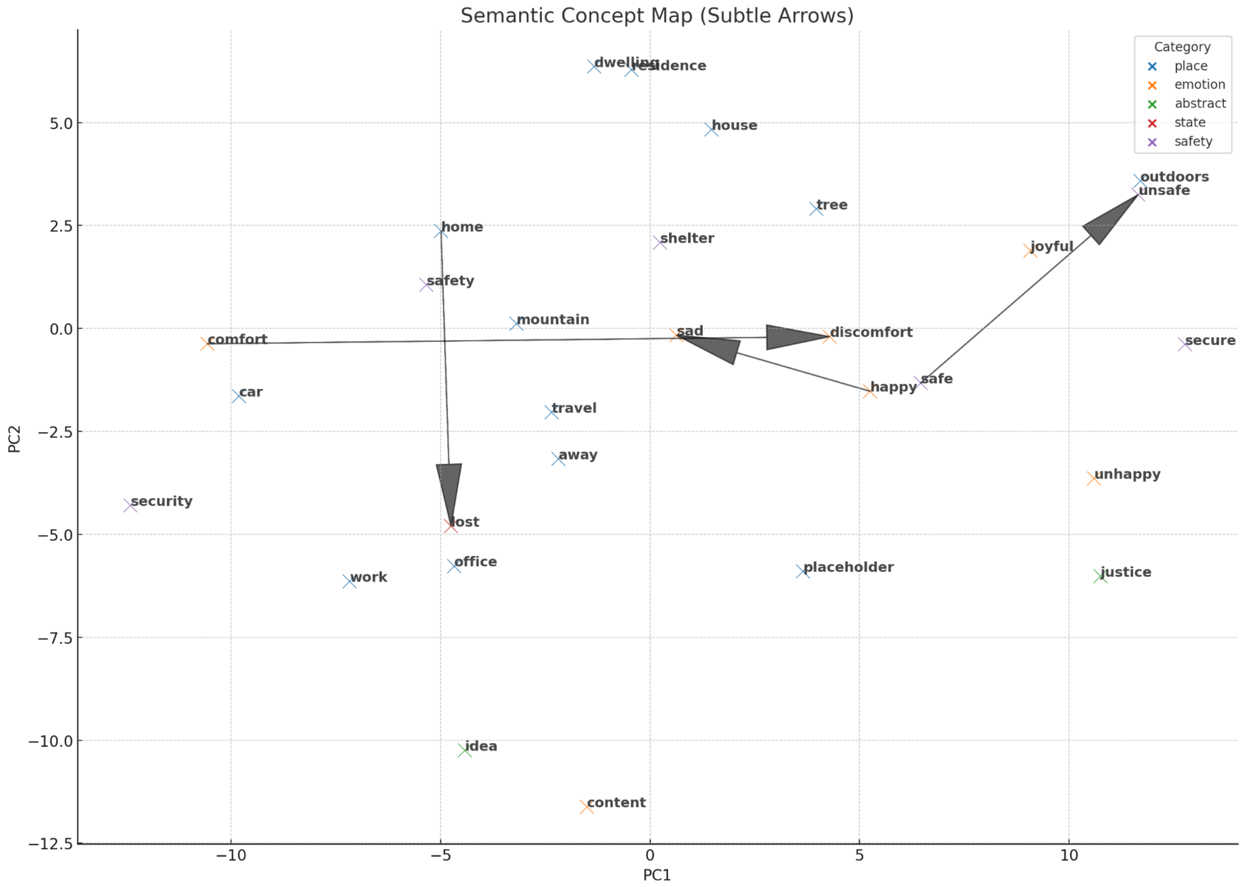
🧠 What the final image shows

The plotted map displays:

- **One point per concept**, representing the centroid of its activation space
- **Color-coded categories** showing emergent clustering (e.g., emotion vs place)
- **Directional arrows** that align with semantic oppositions (e.g., *safe* → *unsafe*)

Despite PCA’s limitations, many concept pairs form clean spatial vectors that are consistent with human interpretations of oppositional meaning.

This provides **empirical support for the linear representation hypothesis** within local concept pairs—while also exposing the fuzzier, drift-prone encoding of more abstract terms.


Semantic Concept Map (Subtle Arrows)
PC1
PC2
−10
−5
0
5
10
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
Category
place
emotion
abstract
state
safety
dwelling
residence
house
outdoors
unsafe
tree
home
shelter
joyful
safety
mountain
sad
discomfort
comfort
secure
safe
happy
car
travel
away
unhappy
security
lost
office
work
placeholder
justice
idea
content

🧭 Semantic Concept Map (Centroid-Only, PCA, Directional Arrows)

A concept-level interpretability experiment on GPT-2 Small

❓What is this?

This is a 2D visualization of how GPT-2 Small encodes a curated set of emotional, spatial, and abstract concepts in its **final hidden layer**. Each labeled point represents the **centroid** of activation vectors for a given concept. Arrows show **semantic directions** derived from linear vector arithmetic.

🛠️ How the experiment was built

Model

- **GPT-2 Small**, 124M parameters, queried via `transformers` locally

Concepts

- 49 curated English-language concepts across 5 high-level semantic domains:
  - `emotion`: e.g., *happy*, *sad*, *content*
  - `safety`: *safe*, *unsafe*, *shelter*
  - `place`: *home*, *house*, *work*, *mountain*
  - `abstract`: *idea*, *justice*
  - `state`: *lost*, *placeholder*, *comparisons*

Prompt Generation

Each concept was embedded in a variety of light-touch sentence templates, including:

- `"I am [concept]"`
- `"They are [concept]"`
- `"This place is [concept]"`
- `"Home is [safe]"`
- `"This is [concept]"`
- `"[concept] was found"`

This was done to test **frame stability** and probe how consistent the internal representation of a concept remains across minimal syntactic/semantic perturbations.

Activation Capture

- From each prompt, we extracted the **final hidden state activation** (layer 12, last token) for the `[concept]` token.
- Each vector is 768-dimensional.
- **~12 prompts per concept**, stored as `.npy` vectors.

Total data collected:

> **49 concepts × 12 prompts = 588 vectors**

Centroid Mapping

- For each concept, we computed the **mean vector (centroid)** across its ~12 prompt activations.
- These 49 centroids were used as the basis for the map—**one point per concept**, simplifying visual interpretation.

Dimensionality Reduction

- **PCA** used to project from 768D to 2D.
- Chosen for its ability to preserve **global geometric structure** and enable **arrow-based reasoning**.
- PCA also allows vector differences (e.g. *safe* → *unsafe*) to be meaningfully visualized.

Categorical Coloring

- Each concept was heuristically categorized into:
  - `emotion`, `safety`, `place`, `abstract`, `state`
- These categories were color-coded on the map for interpretability—not learned or clustered.

Directional Arrows

To test the **Linear Representation Hypothesis**, we manually defined 4 semantic pairings:

- `home → lost`
- `happy → sad`
- `safe → unsafe`
- `comfort → discomfort`

Arrows represent the **literal vector offset in PCA space** between these concept centroids.

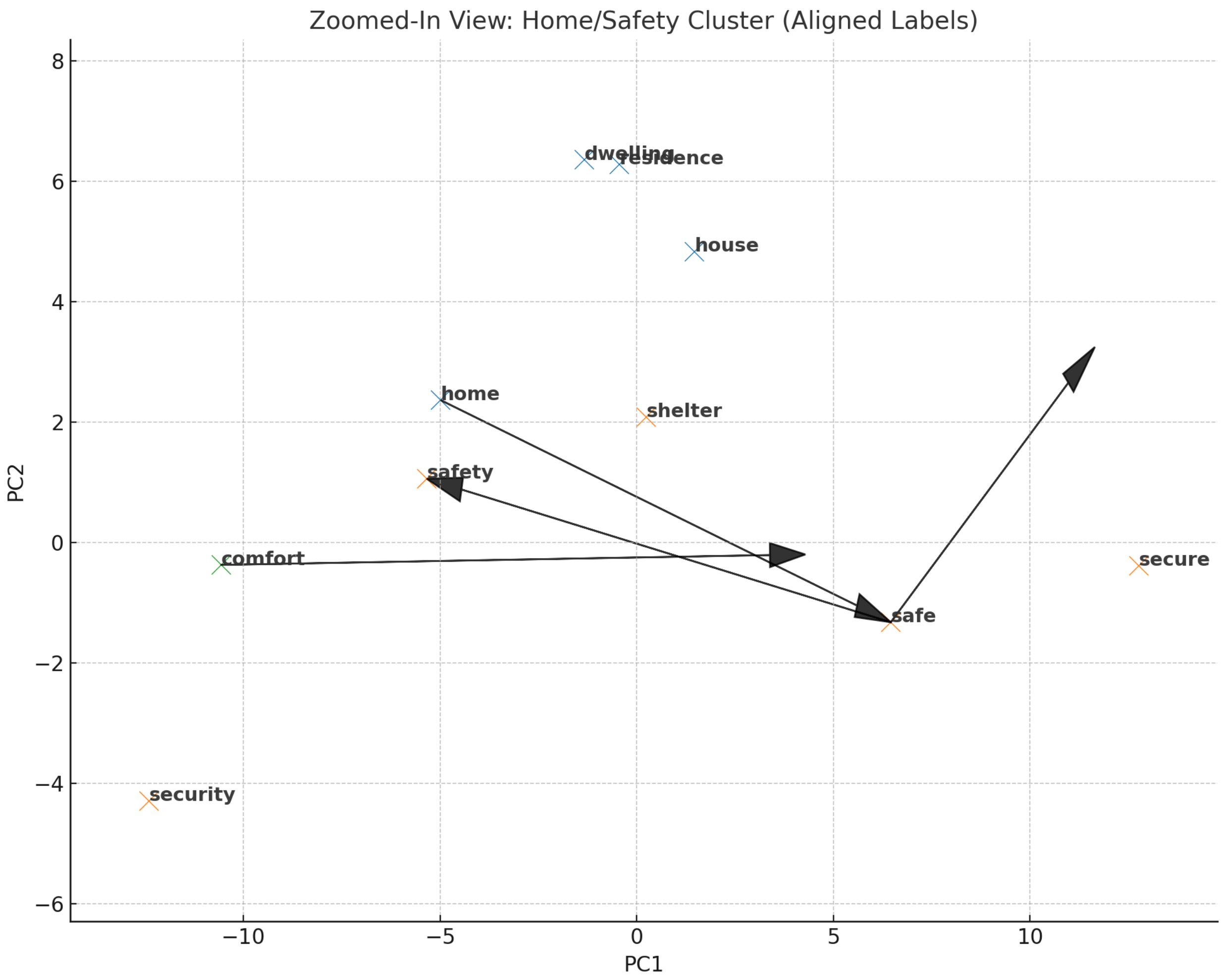
🧠 What does this show?

- **Clustering**: Concepts like *home*, *safety*, *shelter* form intuitive clusters.
- **Directional geometry**: Pairs like *happy* → *sad* and *safe* → *unsafe* yield clean, interpretable directional offsets, validating **linear semantic relations**.
- **Concept drift**: Abstract terms like *justice* and *idea* sit far from dense zones, suggesting the model encodes them in more diffuse or unstable ways.

This provides empirical support—at least in local contexts—for the idea that **conceptual relationships in language models can be spatial, directional, and analogical**.

⚠️ What this is *not*

- Not a claim about truth, ontology, or grounded meaning.
- Not a generalization to larger LLMs or multilingual embeddings.
- Not evaluated with supervised probes—this is zero-shot, unsupervised spatial mapping.
- Not a benchmark, but a **tool for interpretability and emergence detection**.


Zoomed-In View: Home/Safety Cluster (Aligned Labels)
dwelling
residence
house
home
shelter
safety
comfort
secure
safe
security
PC1
PC2
8
6
4
2
0
−2
−4
−6
−10
−5
0
5
10

🔍 Zoomed Concept Map: Home / Safety Cluster (Centroids, Raw Points, Directional Arrows)

This visualization presents a **magnified view** of the local conceptual geometry around *home*, *safety*, *shelter*, and related terms within the PCA-projected activation space of GPT-2 Small.

It isolates a semantically coherent region of the broader concept map and reveals both the **intra-cluster structure** and **frame-level stability** of individual concepts.

🔧 What's Shown Here

- **Centroids** (large dots): Each represents the **mean activation vector** (layer 12) for a concept across ~12 prompt variants.
- **Raw activations** (small, translucent points): Show the actual distribution of activations per concept, capturing how consistent a concept's encoding is across different phrasings.
- **Directional arrows**: Vector deltas drawn between selected concept pairs, rendered from one centroid to another. These arrows are literal displacements in PCA space based on final-layer model activations.

📐 Geometry and Insights

- Concepts like `home`, `safe`, `safety`, `shelter`, and `secure` form a **tight spatial cluster**, suggesting GPT-2 Small encodes these ideas with similar semantic structure.
- The **raw activation points** for many of these (e.g., `safe`, `shelter`) cluster closely around their centroid, indicating **high frame stability**—i.e., the concept's meaning doesn't shift much across prompt variations.
- Other concepts, such as `comfort` or `home`, show more variance, implying **greater context sensitivity** or polysemy.
- Arrows like `comfort → discomfort` or `safe → unsafe` show clean, nearly linear displacements, supporting the **linear representation hypothesis** for antonymic pairs in this subregion.

🧪 Notes on Methodology (specific to this slice)

- The axis limits were manually set to **zoom in on the centroid region**, providing true magnification rather than just filtering.

- Jitter was **not applied** to centroids in the final version—labels are aligned precisely with their representative point.

- The visual density of raw points offers a **proxy for conceptual consistency**: the tighter the spread, the more stable the encoding across linguistic frames.

Simple explanation:

🧭 What is this?

This is a simplified visualisation of how a small AI model represents the meaning of certain words—especially words related to *home*, *safety*, and *comfort*.

It doesn't show how the model *defines* these words, or what it "feels" about them. Instead, it shows how the model stores and positions them inside its own internal structure, based on how it has learned to use language.

🧪 How we made it

We gave the model short example sentences using words like *home*, *safe*, *lost*, *shelter*, etc.

Examples include:

- “I am home”
- “This place is safe”
- “Comfort was found”

The model processes each of these and produces a long list of numbers that represent its internal state when it sees each word. These numbers are called **activations**—you can think of them as a kind of “snapshot” of what the model is doing internally when it hears that word in context.

Each word was tested across a dozen different sentence variations to get a more stable reading. Then, we averaged the results for each word.

Because those activations are high-dimensional (768 numbers long), we used a technique called **PCA** to shrink them down to 2D so we could make a simple map.

🔍 What the map shows

- **Each dot** represents a word like *home*, *safe*, or *shelter*.
- The **location** of each dot shows how the model “places” that concept in relation to others.
- **Arrows** show direct comparisons—for example, from *safe* to *unsafe*, or *comfort* to *discomfort*.
- Smaller, semi-transparent dots show the individual results from each sentence version we used to test the concept.

When we zoom in on this part of the map, we can see that words like *home*, *safety*, and *shelter* cluster quite closely together. This suggests the model treats these ideas as similar—at least within the scope of these simple prompts.

Some words—like *safe*—produce very consistent results across different sentences. Others—like *comfort*—show more spread, which might mean they’re more sensitive to phrasing or context.

🤔 What this does (and doesn’t) mean

This doesn’t tell us what “home” means to the model in a human sense. It doesn’t reveal a deep understanding. But it does let us see **how consistent and connected** certain concepts are in the model’s internal space.

It’s a way of observing patterns in how a model represents language—not in terms of definitions, but in terms of structure and proximity.

That’s the basic idea behind this kind of experiment:
We’re trying to look at meaning *as shape*—and see what kinds of patterns show up.