**VERSION 9 Basic Documentation**

Despite the later version name this software is older than the 7.5xx suite. It was used earlier on for more general neuron probing and analysis.

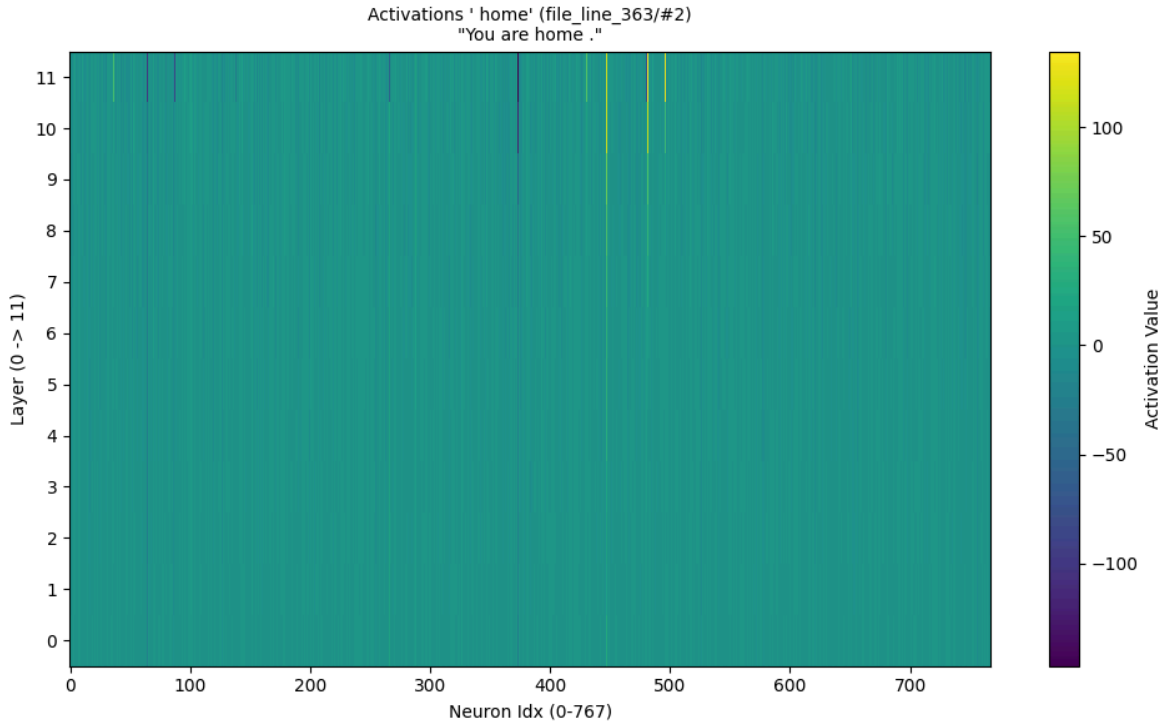
Inside the zipped full folder is the critical code for reproducibility, explained below, as well as copies of my results, and saved chatlogs between myself and AI discussing, analysing, and reflecting on it all.

---

**1.** `analyze_batch_v9.0.py`
(*This is the big one — the brain of the v9.0 phase*)

- Loads **GPT-2 Small** through TransformerLens.
- Feeds it structured prompts one by one.
- Captures the **activations** (specifically *MLP post-activations*) across all layers at a specific token.
- Saves:
    - The raw activations (`.npy` files).
    - Optional heatmap images of activation vectors (`.png`).
- Analyzes:
    - Which neurons activate most for different concepts.
    - Collapse events (repetitive or incoherent outputs).
    - Comparative differences between concept activations (if comparisons are configured).
    - Vector operations (arithmetic on neuron activations, if configured).
- Optionally calculates a **Conceptual Drift Index (CDI)** to quantify how "far" activations drift for a given concept.

- Optionally generates a **2D visualization** (using UMAP or t-SNE) of final layer activations.




- Produces full and brief summary reports.

**In short:**
This script is a full activation capture + analysis + visualization pipeline aimed at exploratory neuron and concept behavior mapping.
It's heavy-duty and flexible — built for structured batch probing.

---

### 2. `run_probe_batch.py`

(*This is much simpler — a little helper tool*)

- Loads **GPT-2 Small** directly via Huggingface.
- Reads a list of prompts from a file (`drift_probe_prompts_v1.txt`).
- Runs each prompt through GPT-2 Small in inference mode.
- Saves the *prompt* and *model response* to a `.jsonl` file (`probe_outputs_v1.jsonl`).

**In short:**

This is just a **prompt-response logger**.

You use it to quickly gather model outputs for various prompts — likely to find examples of collapse, weirdness, or drift before doing heavier neuron analysis.