**Question (general):** Is the annotation CMV enhancer represented in the model hidden states of gLMs?

**General set-up across tasks:**Trained linear probes on MLP-activation vectors from layer 12 of the nucleotide transformer (NT-50m v2). Used multiple random seeds (5 runs with different initializations).

1. **Original Task (Baseline)**

* Labels: has annotation ‘CMV enhancer’ yes/no?
* Achieved the highest performance with F1 score of 0.717 ± 0.022
* Used class-weighted loss to handle imbalanced data

1. **Control Task: Random Labels**

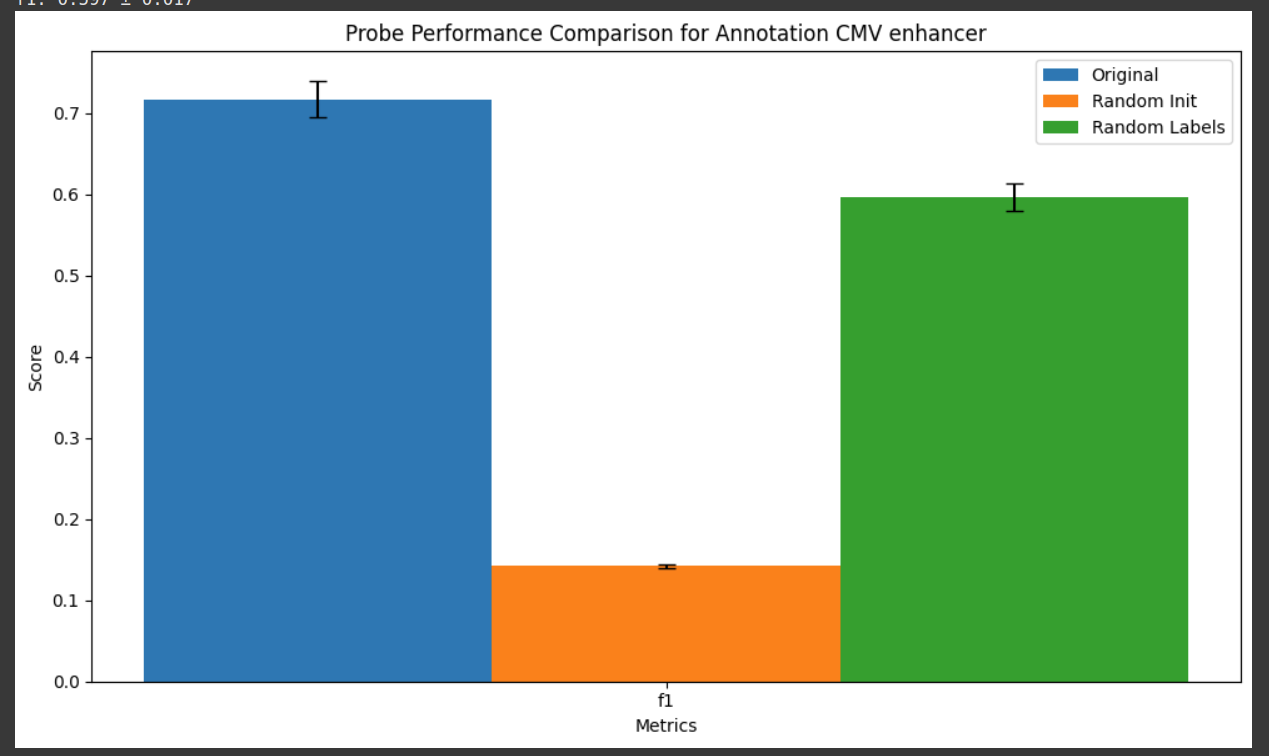
* Kept the same input representations but randomly shuffled the labels
* Maintained exact class distribution and token consistency (same sequence always got same label)
* Achieved F1 score of 0.597 ± 0.017
* This control tests probe's ability to memorize arbitrary mappings

1. **Control Task: Random Initialization**

* Used a randomly initialized transformer model instead of the trained one
* Kept original labels and class distributions
* Achieved lowest F1 score of 0.142 ± 0.003
* Tests whether learned representations are meaningful compared to random ones

**Key Findings:**

* The original task significantly outperformed both control tasks
* The gap between original and random labels (selectivity ≈ 0.12) indicates the probe is capturing real patterns rather than just memorizing
* The large gap between original and random initialization (≈ 0.58) suggests the transformer has learned meaningful representations during pretraining



**Code:** [**Training Linear Probes on NT.ipynb**](https://colab.research.google.com/drive/1to8Y-2A0RGPTtTucT_czXVZLjdW-WpOu?usp=sharing)