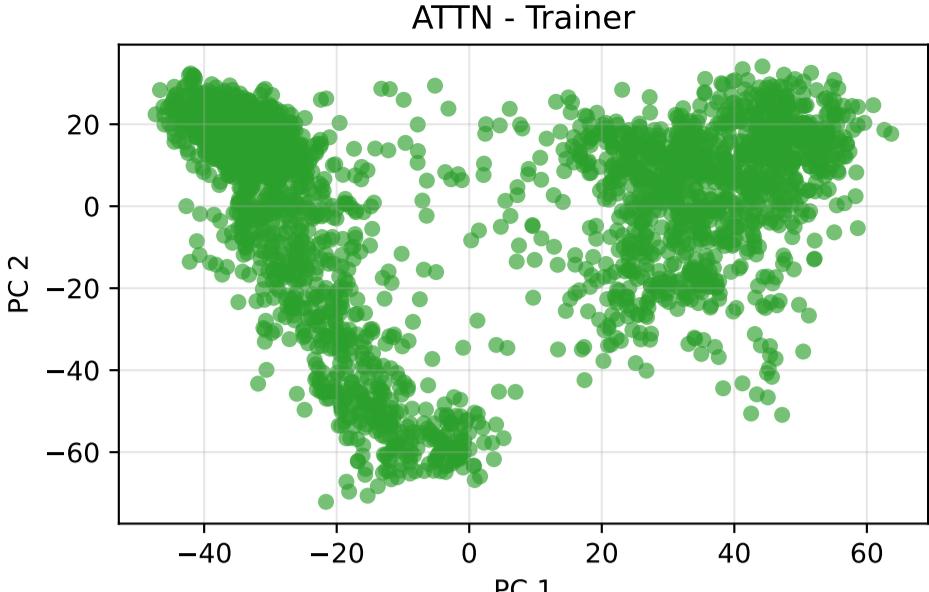
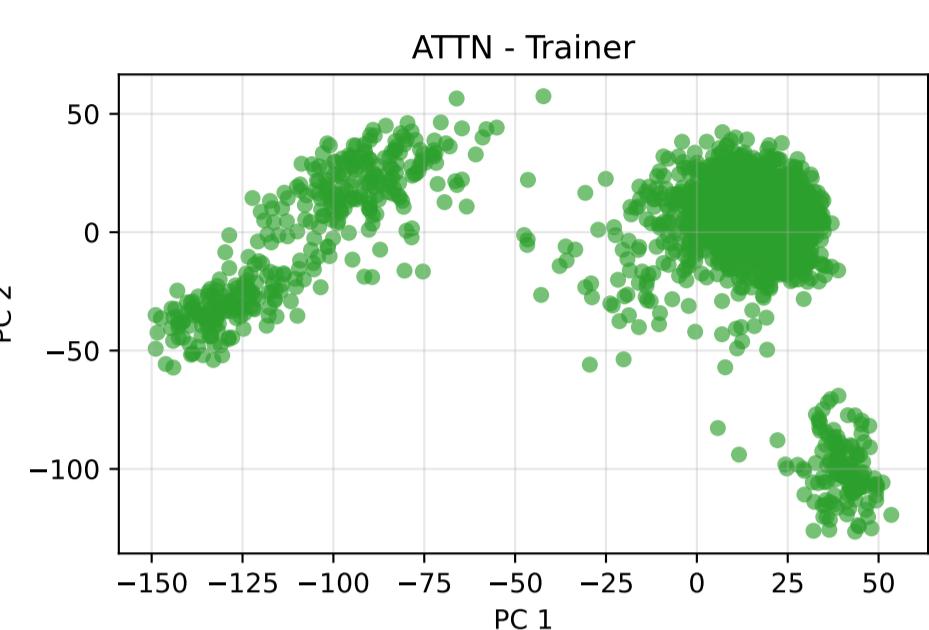


**Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it**

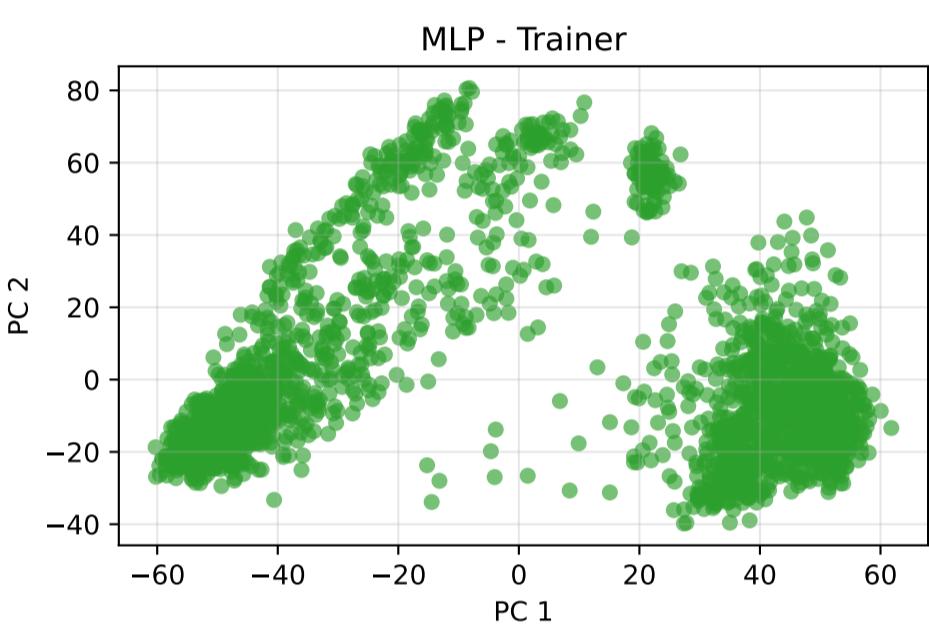


**Trainer: gemma-2-9b-it
Tester: Llama-3 1-8B-Instruct**



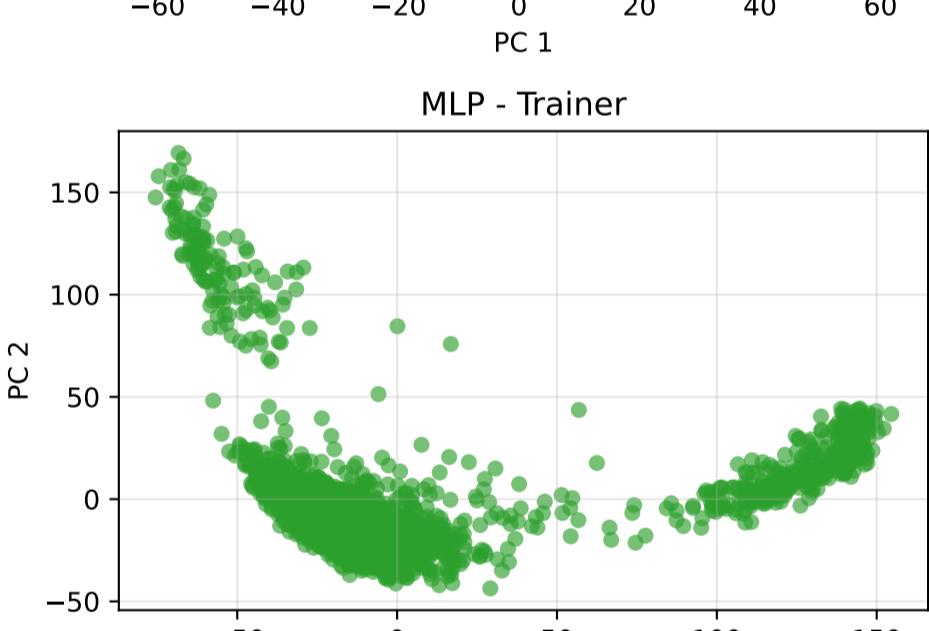
A scatter plot titled "ATTN - Tester pre-align" showing data points in a 2D space defined by PC 1 (x-axis) and PC 2 (y-axis). The x-axis ranges from approximately -50 to 60, and the y-axis ranges from -60 to 25. The data points are colored orange and form several distinct clusters. There are two large, roughly symmetric clusters centered around PC 1 values of -35 and 35, respectively. Between these major clusters, there are three smaller, more vertically oriented clusters located near PC 1 values of -10, 0, and 10. A horizontal grid is present at each integer value from -60 to 25.

**Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it**



A PCA plot titled "MLP - Tester pre-align". The x-axis is labeled "PC 1" and ranges from approximately -80 to 160. The y-axis is labeled "PC 2" and ranges from -50 to 170. The plot shows a complex, non-linear distribution of orange circular data points. The points are densely clustered in several distinct regions: a large upper-left cluster between PC 1 ≈ -80 and -20 and PC 2 ≈ 50 and 160; a lower-left cluster between PC 1 ≈ -80 and 0 and PC 2 ≈ -50 and 50; a central cluster between PC 1 ≈ -20 and 50 and PC 2 ≈ -50 and 50; and a lower-right cluster between PC 1 ≈ 50 and 160 and PC 2 ≈ -50 and 50. There are also several isolated points scattered outside these main clusters.

**Trainer: gemma-2-9b-it
Tester: Llama-3.1-8B-Instruct**



A scatter plot titled "MLP - Tester pre-align" showing data points in PC 1 vs PC 2 space. The x-axis is labeled "PC 1" and ranges from -50 to 150. The y-axis is labeled "PC 2" and ranges from -40 to 80. The data points are orange circles, forming several distinct clusters. One large cluster is centered around PC 1 ≈ 20 and PC 2 ≈ 40. Another large cluster is centered around PC 1 ≈ 100 and PC 2 ≈ 0. There are also smaller clusters at PC 1 ≈ -10, PC 2 ≈ -20; PC 1 ≈ 50, PC 2 ≈ -30; and PC 1 ≈ 100, PC 2 ≈ -40.

MLP - Trainer vs Tester post-align

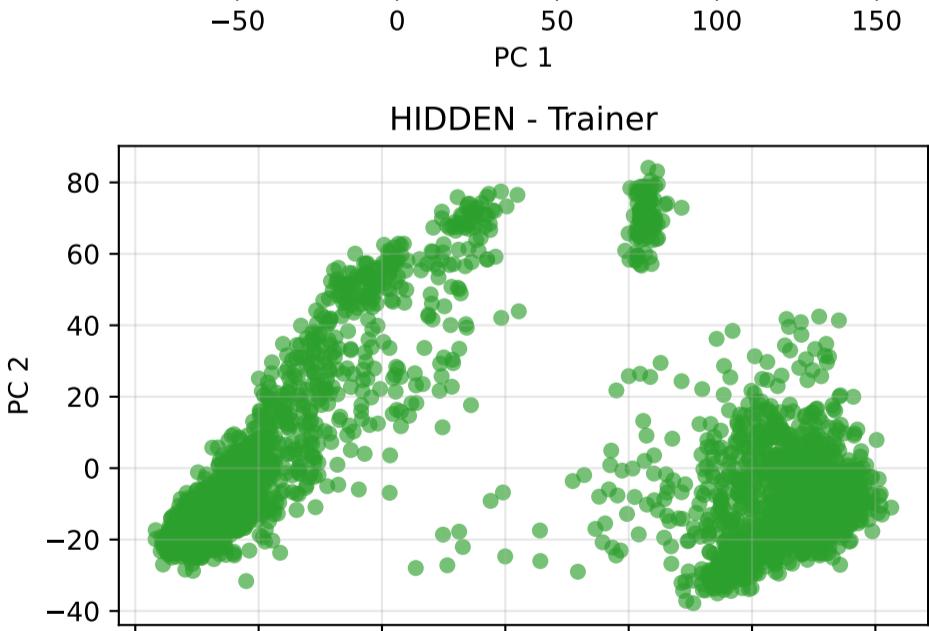
PC 1

PC 2

Trainer

Tester post-align

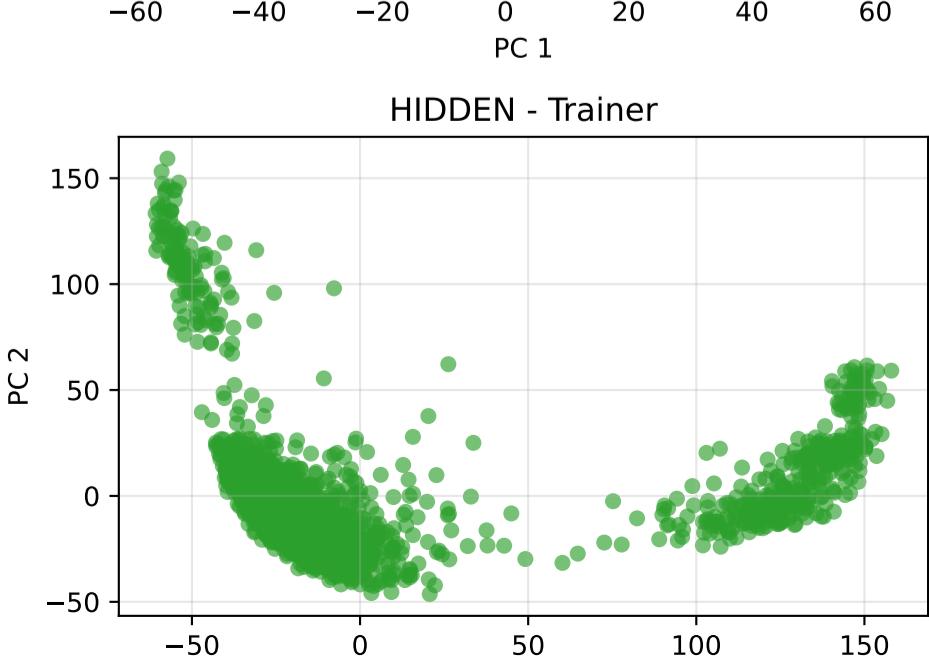
**Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it**



A PCA plot titled "HIDDEN - Tester pre-align". The x-axis is labeled "PC 1" and ranges from -60 to 60. The y-axis is labeled "PC 2" and ranges from -50 to 150. The plot shows a dense cluster of orange data points forming a complex, elongated shape. The points are concentrated along a diagonal line from approximately (-55, 150) to (55, -50), with a significant concentration near the origin (0,0). There are also several outliers and smaller clusters of points extending from the main body.

HIDDEN - Trainer vs Tester post-align

**Trainer: gemma-2-9b-it
Tester: Llama-3.1-8B-Instruct**



HIDDEN - Trainer vs Tester post-align

PC 1

PC 2

Trainer

Tester post-align