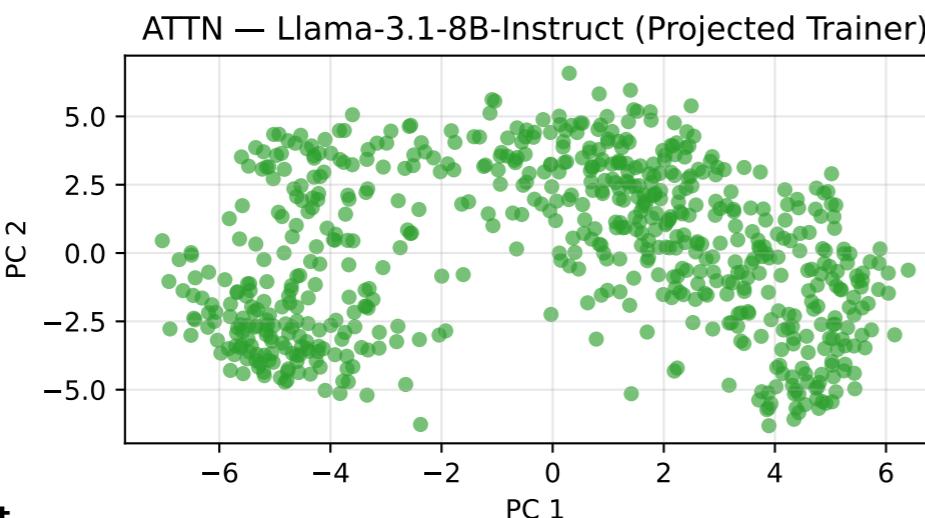
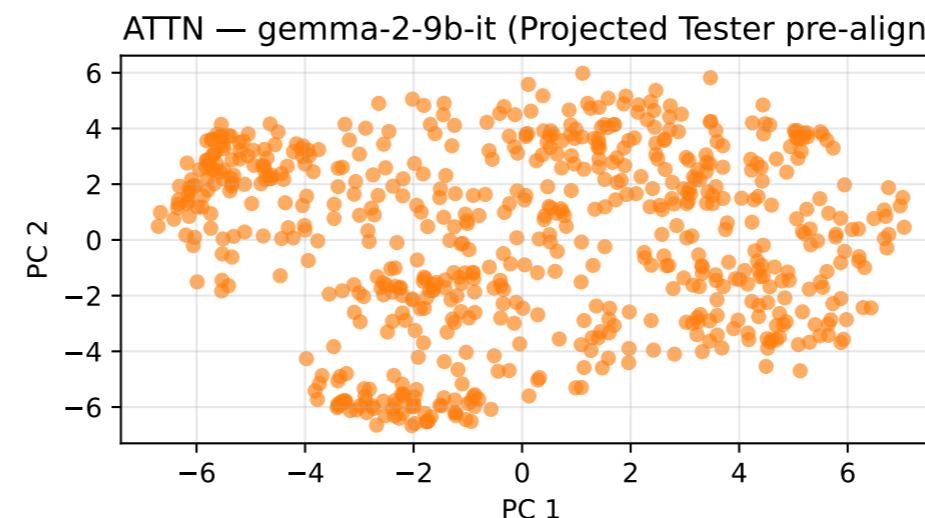


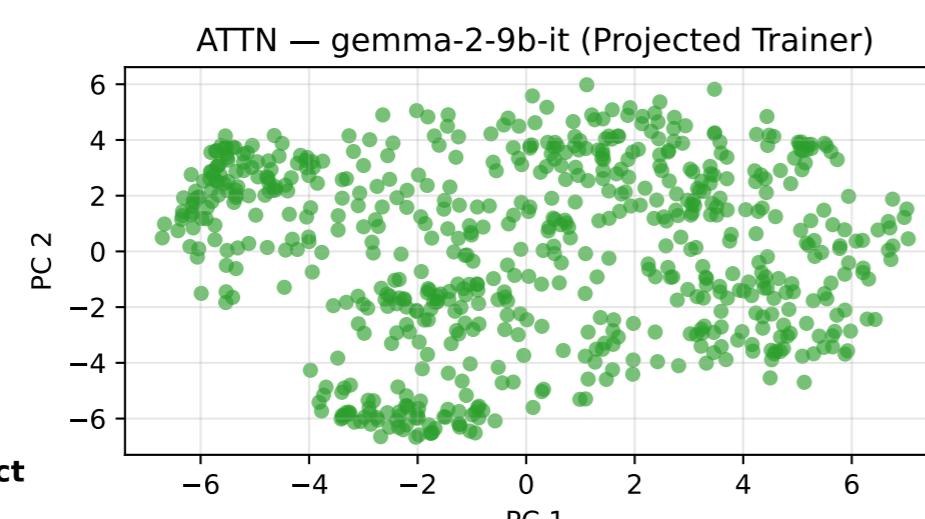
**Trainer: Llama-3.1-8B-Instruct  
Tester: gemma-2-9b-it**



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



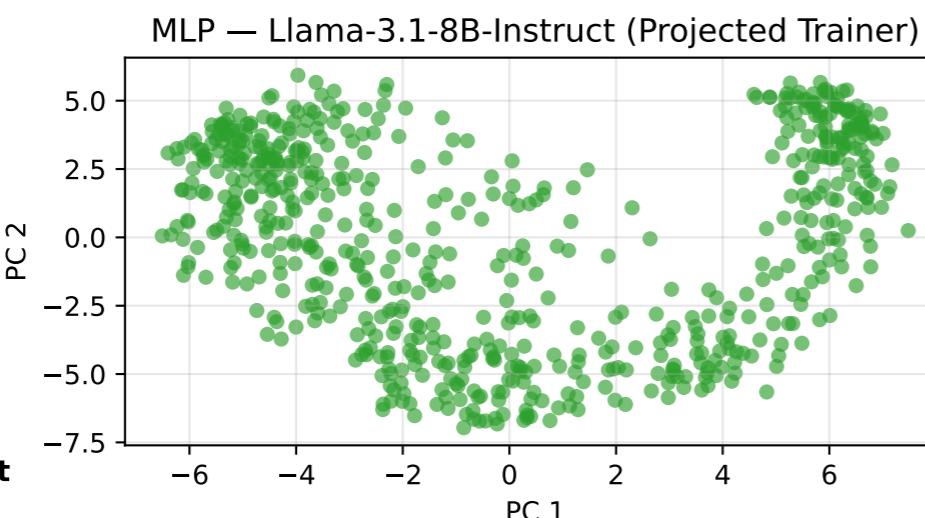
**Trainer: Llama-3.1-8B-Instruct  
Tester: gemma-2-9b-it**



) ATTN — Projected Trainer vs Projected Tester post-align

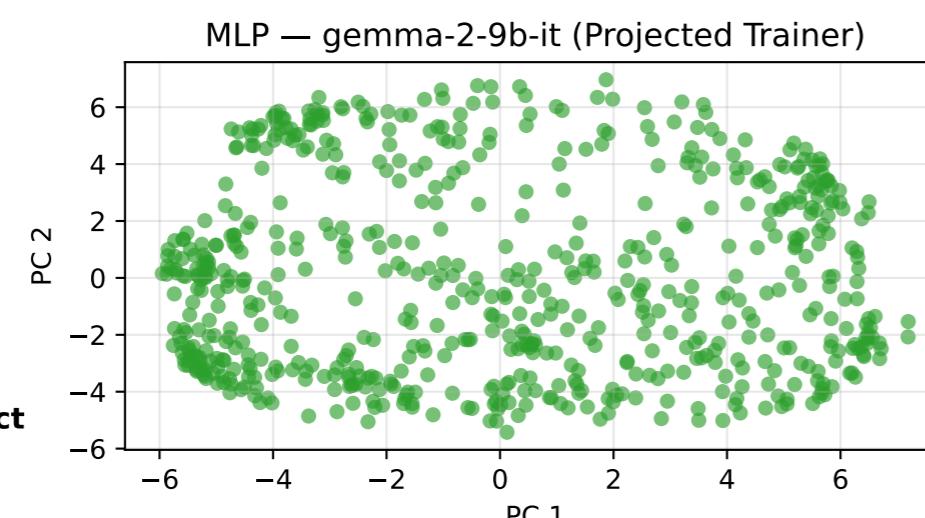
This scatter plot visualizes the alignment of projected trainer and tester embeddings after post-alignment. The x-axis is labeled "PC 1" and the y-axis is labeled "PC 2". The plot shows two main clusters of points: a large cluster of green dots representing "Projected Trainer" and a smaller, distinct cluster of orange dots representing "Projected Tester post-align". The orange cluster is centered around PC 1 values between -1 and 1 and PC 2 values between -1 and 1. The green cluster is more spread out, with many points extending from PC 1 = -6 to 6 and PC 2 = -6 to 6. A legend in the bottom right corner identifies the two series.

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



A scatter plot titled "MLP — gemma-2-9b-it (Projected Tester pre-align)". The x-axis is labeled "PC 1" and ranges from -6 to 6. The y-axis is labeled "PC 2" and ranges from -6 to 6. The plot shows a complex, non-linearly separable dataset represented by orange circular points. The points are clustered into several distinct regions, indicating multiple classes or states. A light gray grid is visible in the background.

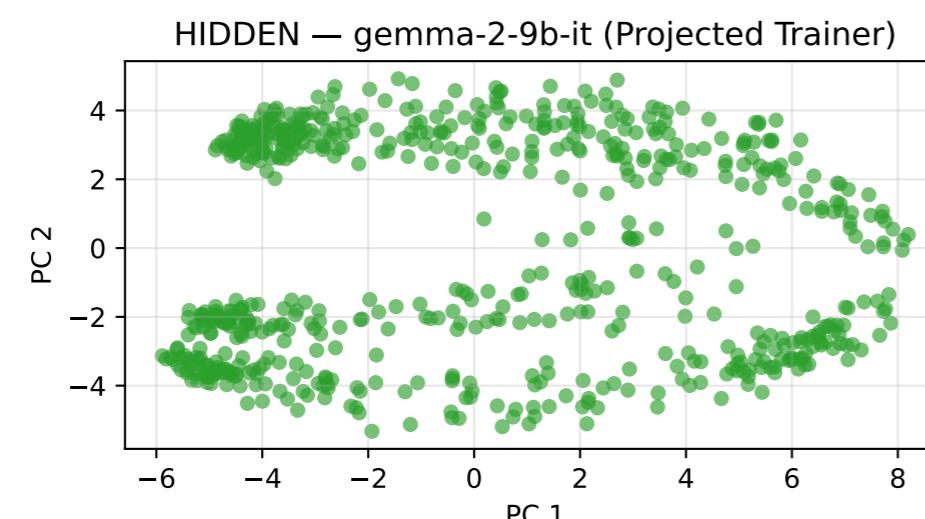
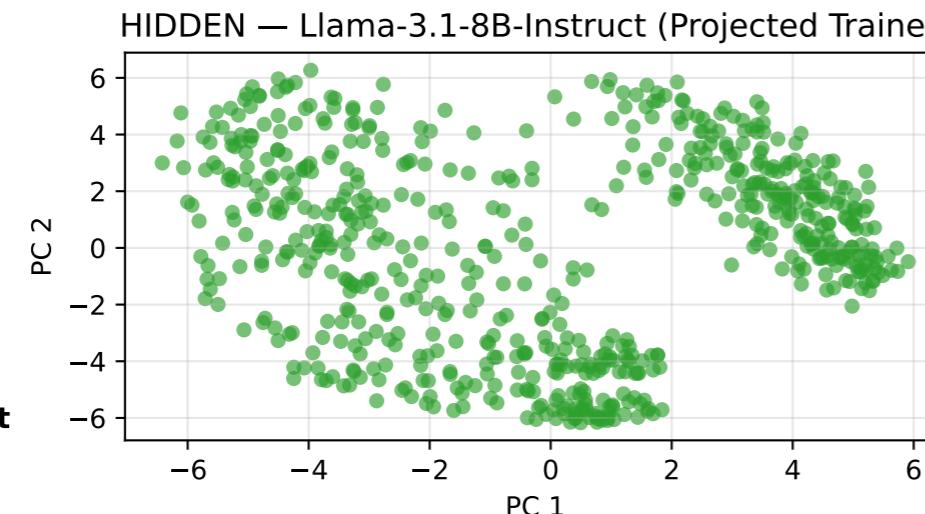
## **Trainer: Llama-3.1-8B-Instruct Tester: gemma-2-9b-it**



A PCA plot showing the distribution of Llama-3.1-8B-Instruct embeddings. The x-axis is labeled "PC 1" and the y-axis is labeled "PC 2". The plot area has a grid with major ticks at -6, -4, -2, 0, 2, 4, 6, and 8 on the x-axis, and -7.5, -5.0, -2.5, 0.0, 2.5, 5.0 on the y-axis. The data points, represented by orange circles, are clustered into several groups. There are two main vertical clusters on the left side of the plot, one centered around PC 1 = -4 and another around PC 1 = 6. Between these, there are several smaller, more horizontal clusters. A notable feature is a vertical column of points centered around PC 1 ≈ 0.5, extending from PC 2 ≈ -6 up to PC 2 ≈ 5. Another vertical column is visible near PC 1 ≈ 2.5, ranging from PC 2 ≈ -6 to PC 2 ≈ 2.5.

Scatter plot titled "MLP — Projected Trainer vs Projected Tester post-align" showing PC 2 vs PC 1. The plot features two clusters: "Projected Trainer" (green circles) and "Projected Tester post-align" (orange circles). A legend in the top right identifies the two series. The axes range from -6 to 6 for both PC 1 and PC 2.

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



A PCA plot showing the distribution of Llama-3.1-8B-Instruct embeddings. The x-axis is labeled "PC 1" and ranges from -6 to 6. The y-axis is labeled "PC 2" and ranges from -6 to 6. The plot area is a grid with major lines every 2 units. Numerous orange circular points represent individual embeddings, forming several distinct clusters. A large central cluster is centered around (-3, 0). Other clusters are located at approximately (-5, 5), (0, -5), (2, 5), (4, 2), and (5, -1).

n) HIDDEN — Projected Trainer vs Projected Tester post-align

This PCA plot visualizes the relationship between projected trainer and projected tester data points after alignment. The x-axis is labeled "PC 1" and ranges from -6 to 8. The y-axis is labeled "PC 2" and ranges from -4 to 4. A legend in the bottom right corner identifies two data series: "Projected Trainer" represented by green circles, and "Projected Tester post-align" represented by orange circles. The green points are clustered primarily in the upper-left and lower-left regions of the plot, while the orange points are concentrated in the center. A grid of light gray lines is overlaid on the plot area.