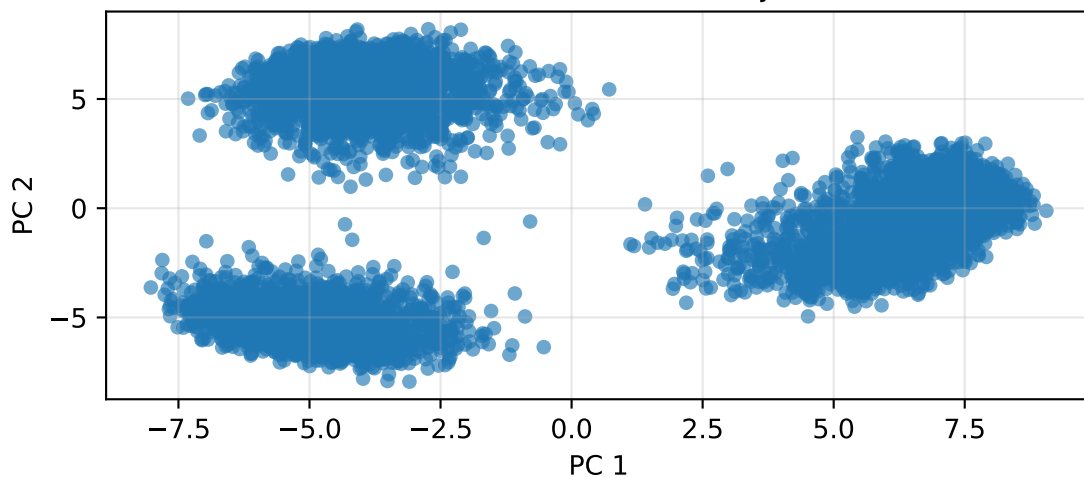
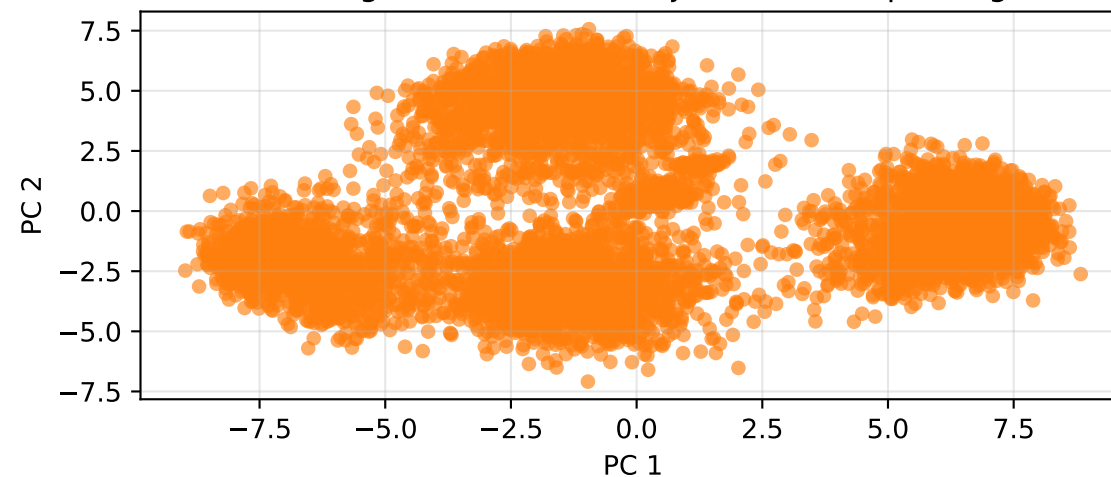


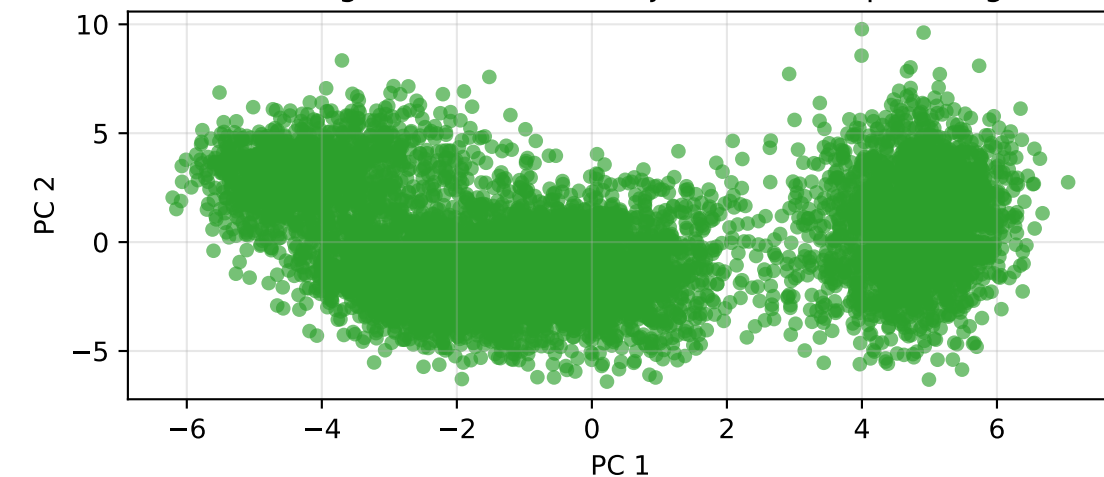
ATTN — Llama-3.1-8B-Instruct (Projected Trainer)



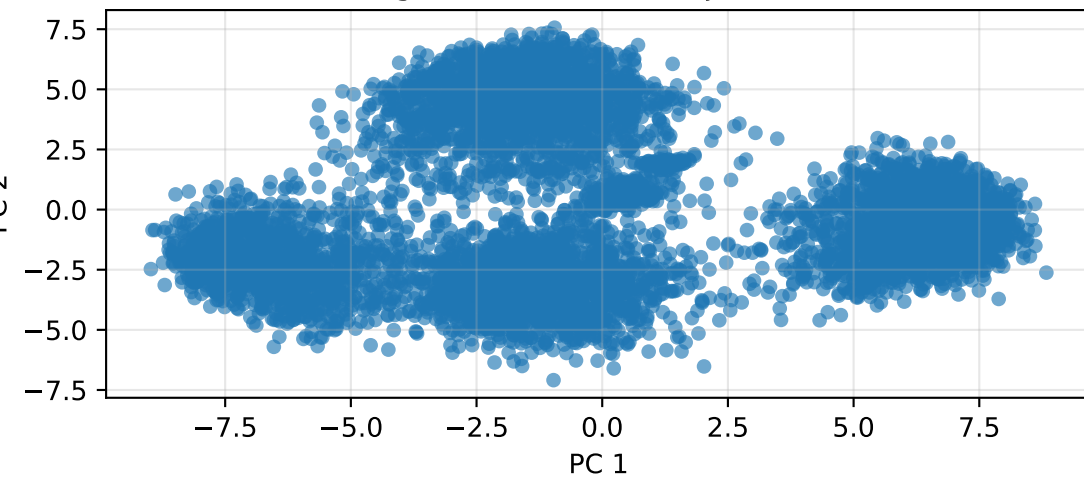
ATTN — gemma-2-9b-it (Projected Tester pre-align)



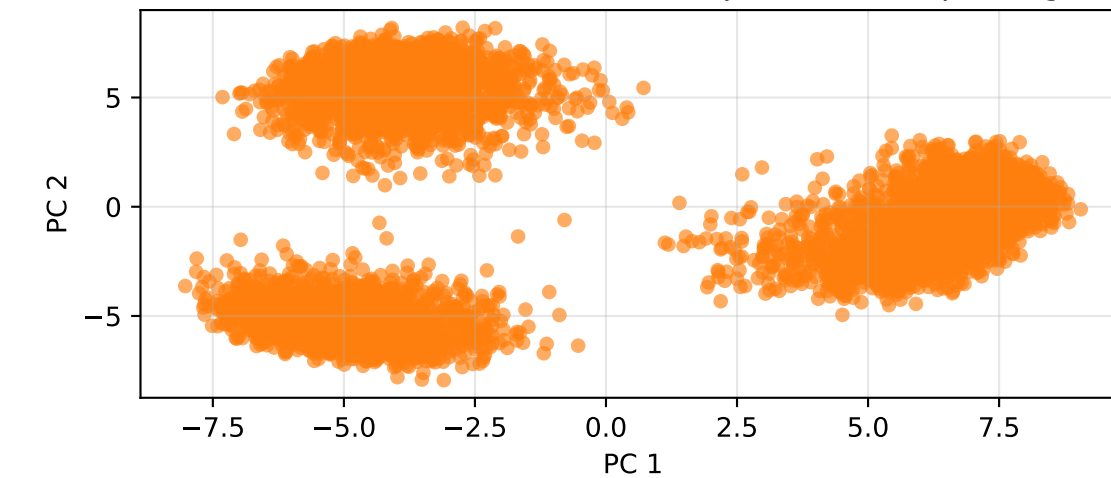
ATTN — gemma-2-9b-it (Projected Tester post-align)



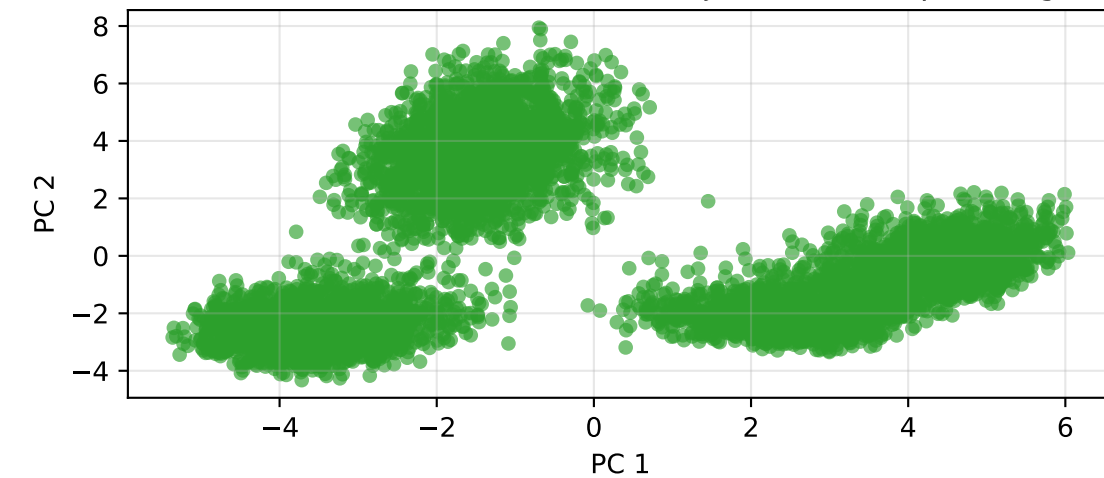
ATTN — gemma-2-9b-it (Projected Trainer)



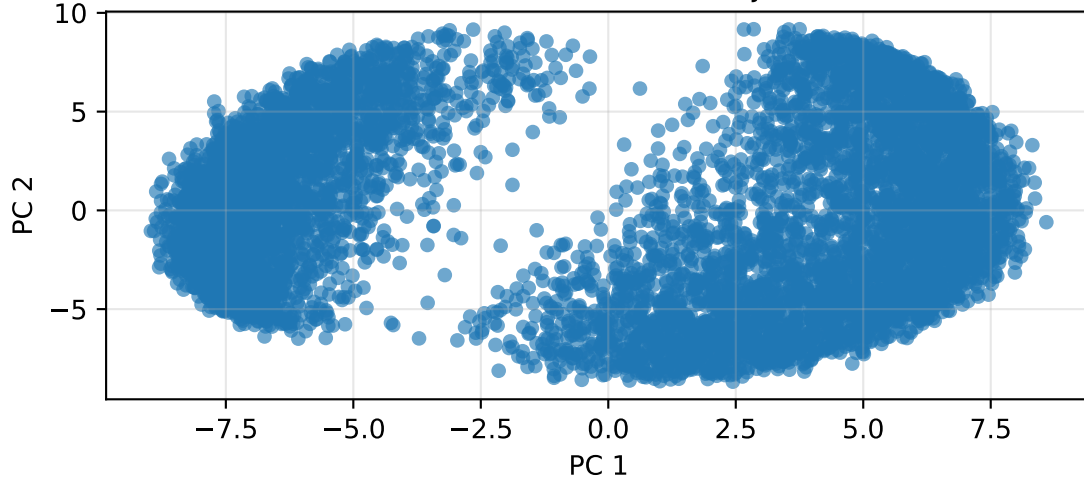
ATTN — Llama-3.1-8B-Instruct (Projected Tester pre-align)



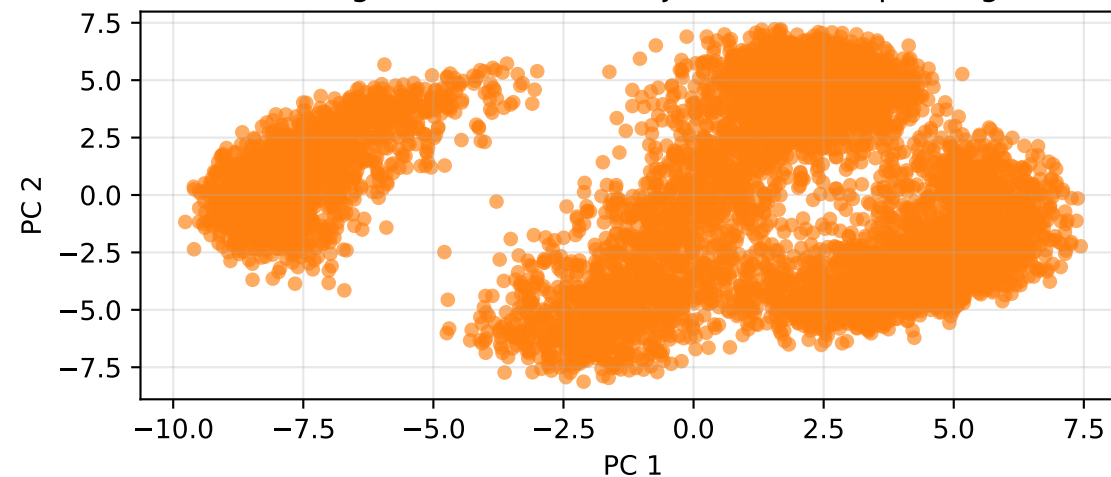
ATTN — Llama-3.1-8B-Instruct (Projected Tester post-align)



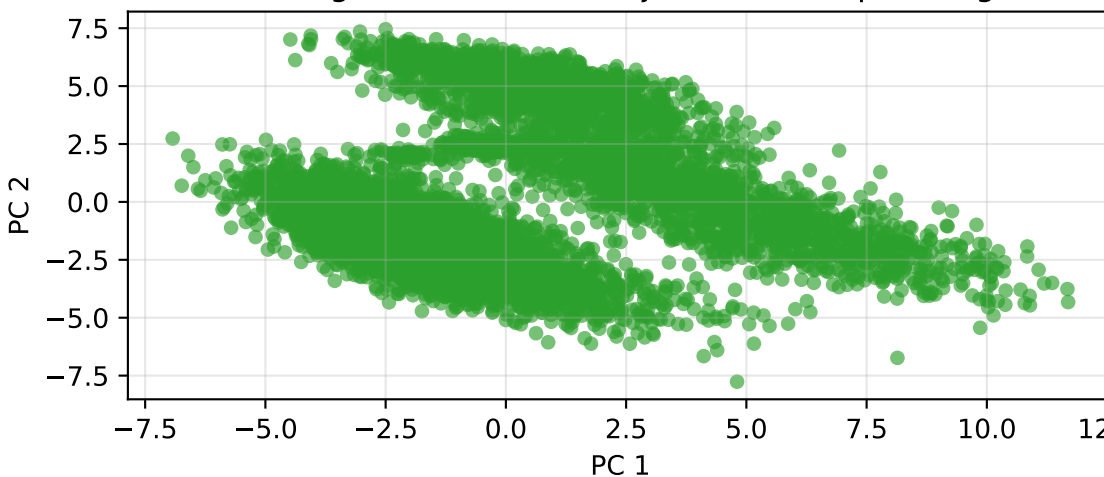
MLP — Llama-3.1-8B-Instruct (Projected Trainer)



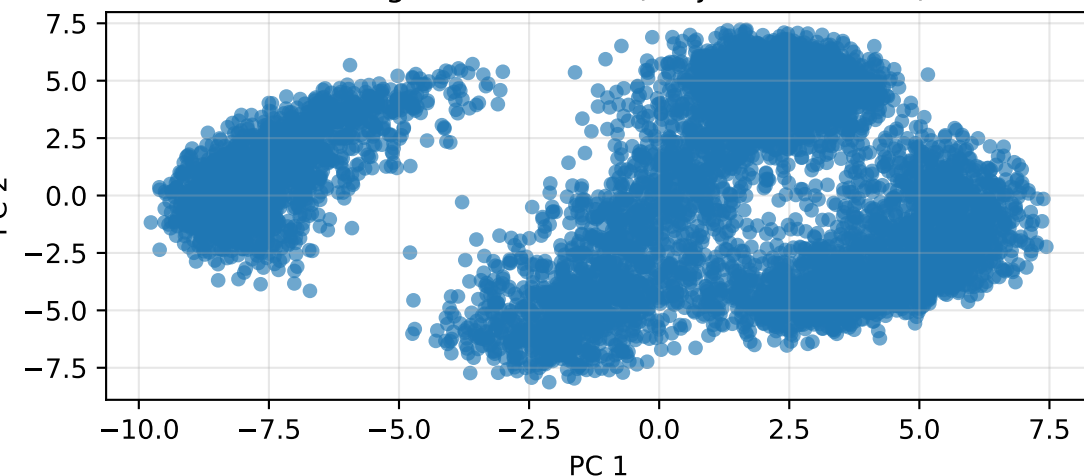
MLP — gemma-2-9b-it (Projected Tester pre-align)



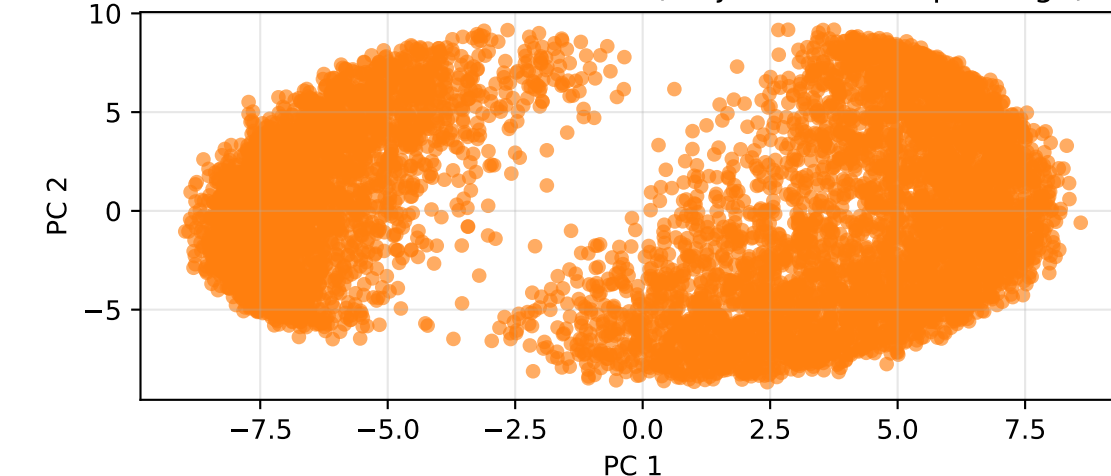
MLP — gemma-2-9b-it (Projected Tester post-align)



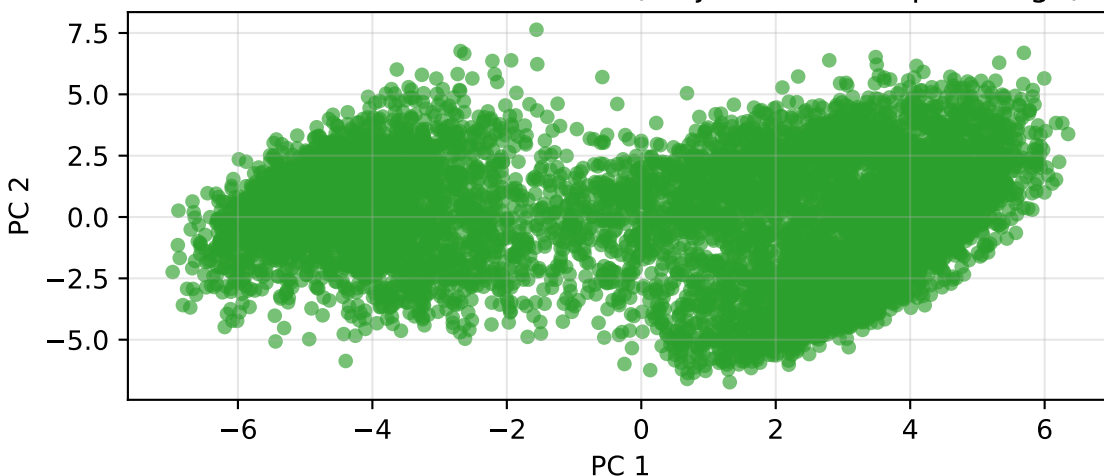
MLP — gemma-2-9b-it (Projected Trainer)



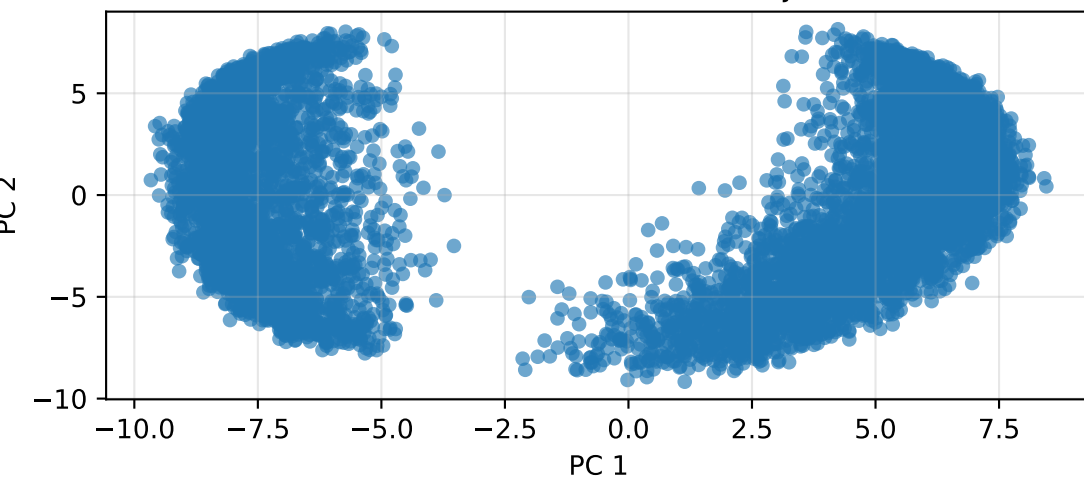
MLP — Llama-3.1-8B-Instruct (Projected Tester pre-align)



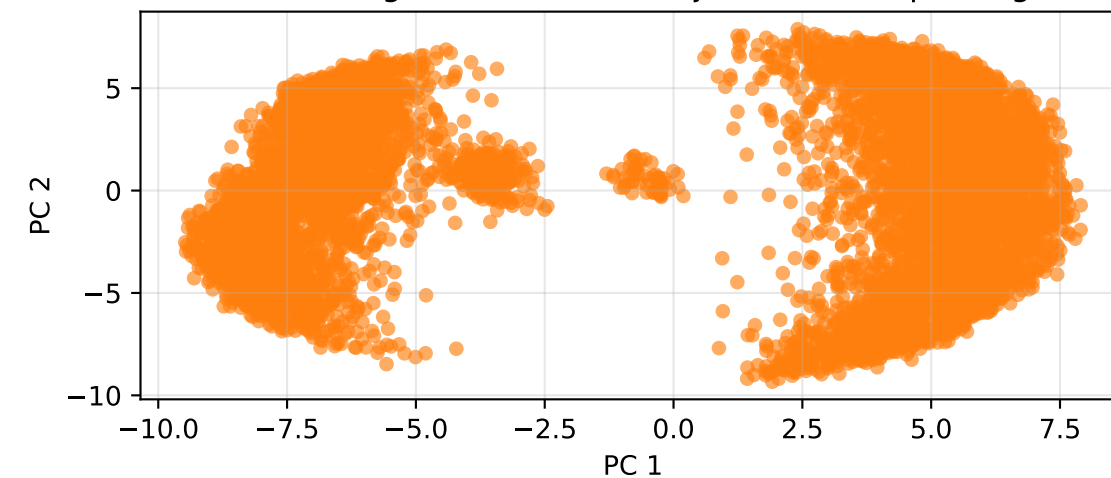
MLP — Llama-3.1-8B-Instruct (Projected Tester post-align)



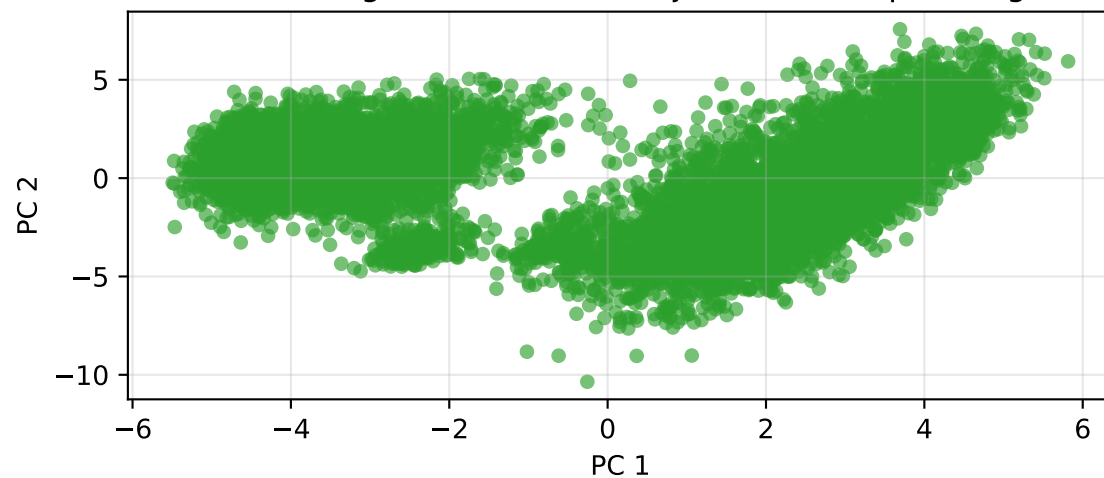
HIDDEN — Llama-3.1-8B-Instruct (Projected Trainer)



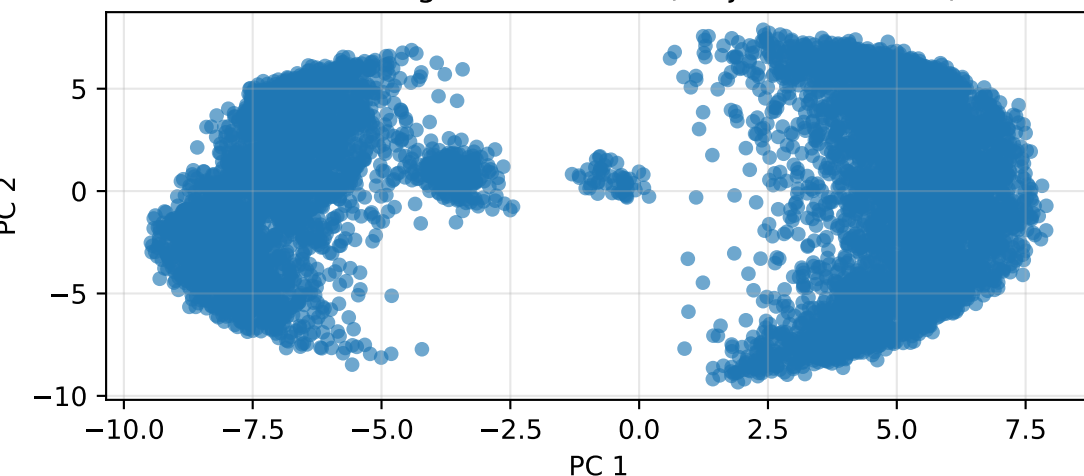
HIDDEN — gemma-2-9b-it (Projected Tester pre-align)



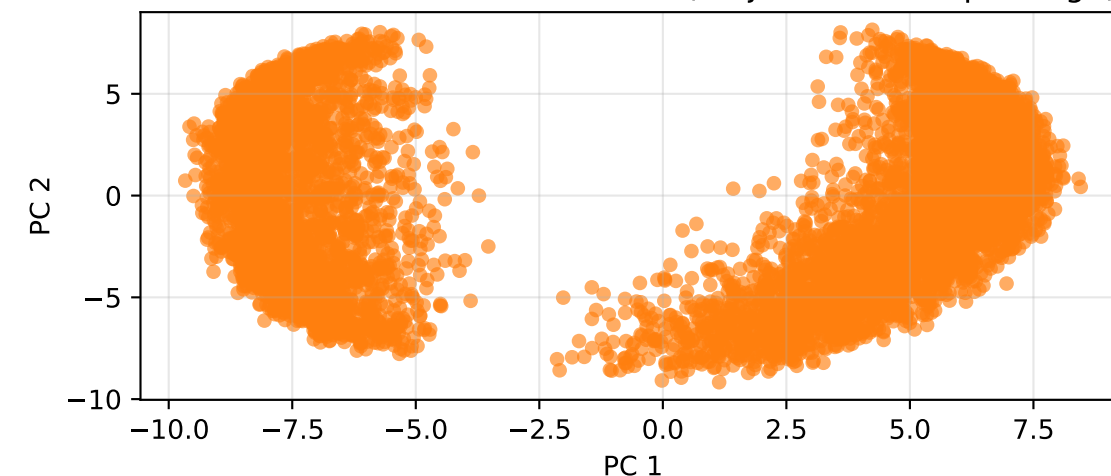
HIDDEN — gemma-2-9b-it (Projected Tester post-align)



HIDDEN — gemma-2-9b-it (Projected Trainer)



HIDDEN — Llama-3.1-8B-Instruct (Projected Tester pre-align)



HIDDEN — Llama-3.1-8B-Instruct (Projected Tester post-align)

