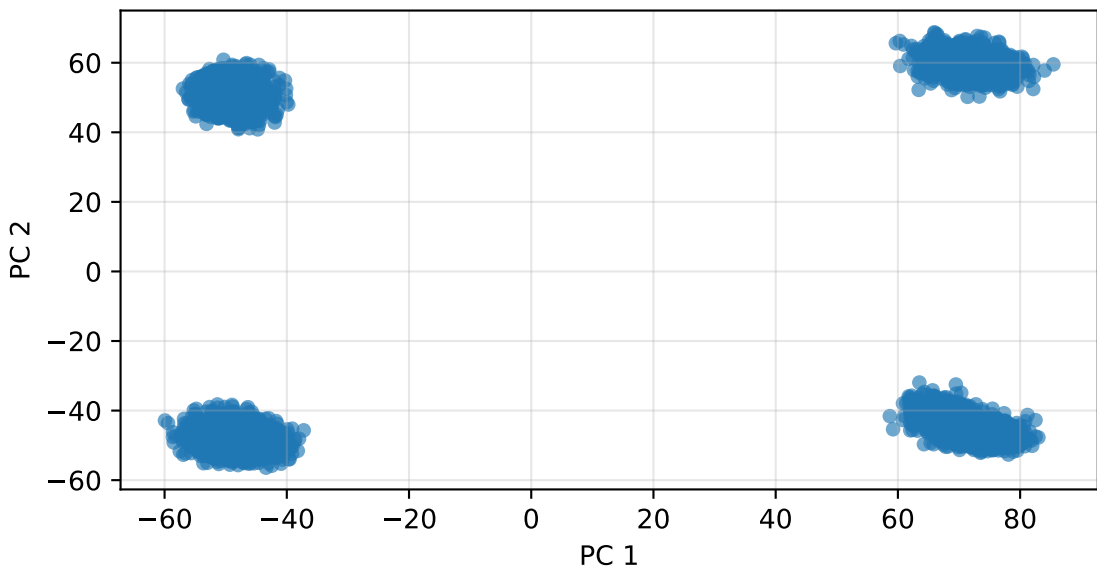
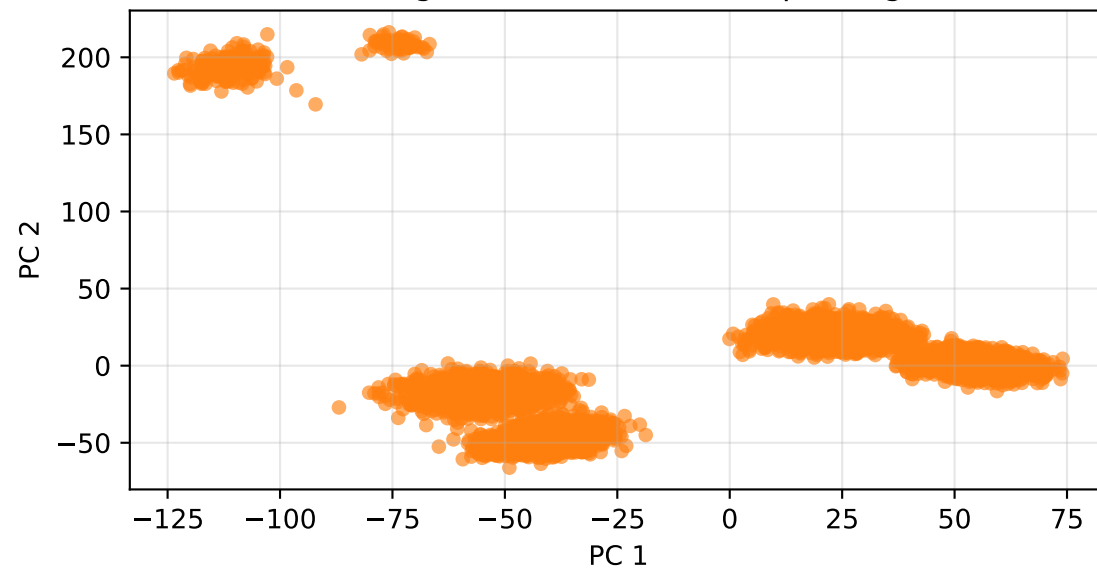


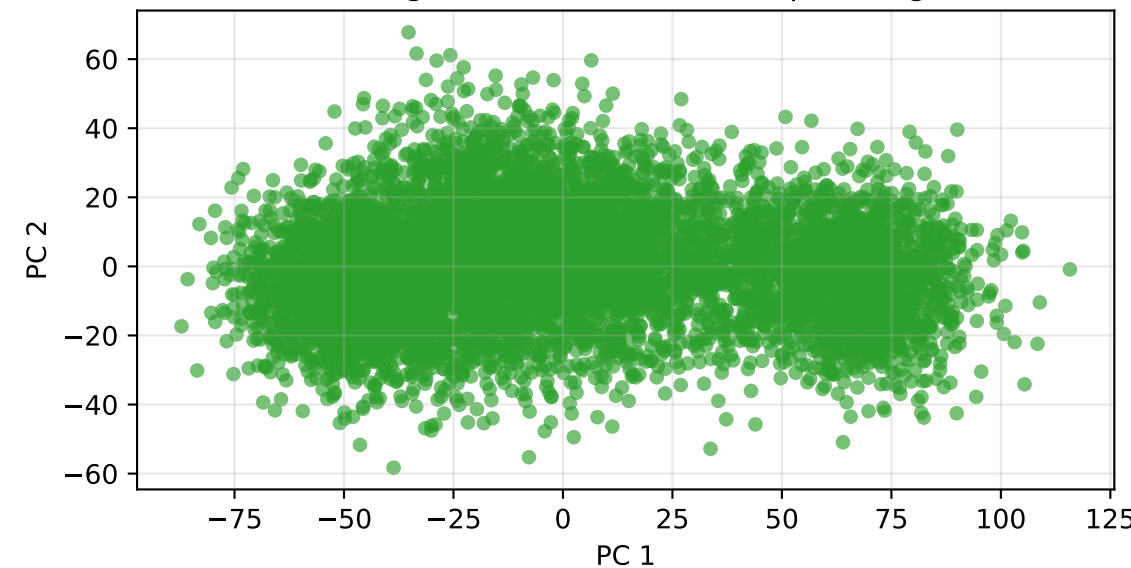
ATTN - Llama-3.1-8B-Instruct (Trainer)



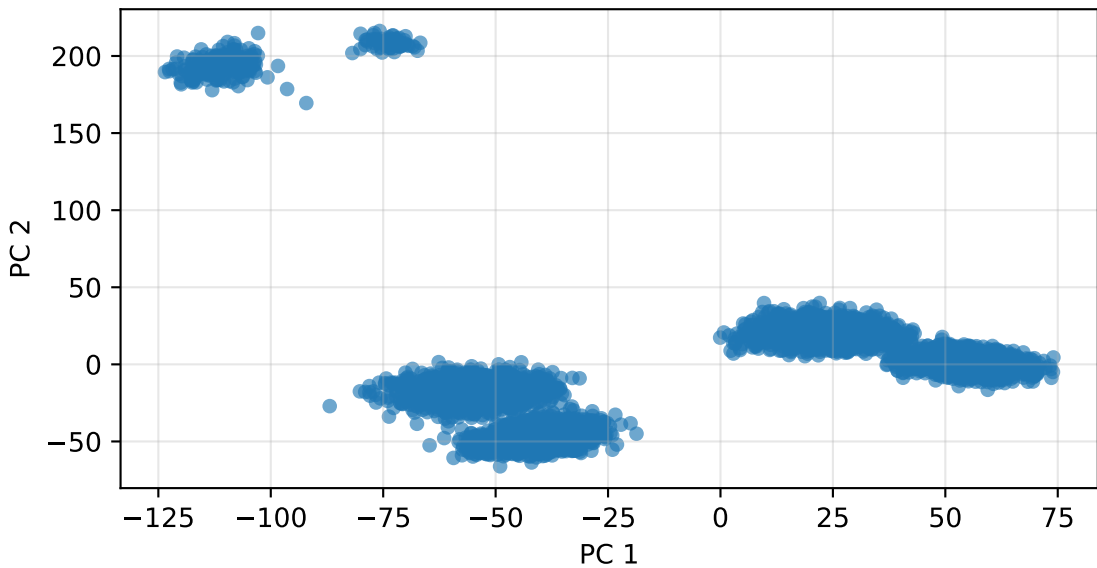
ATTN - gemma-2-9b-it (Tester pre-align)



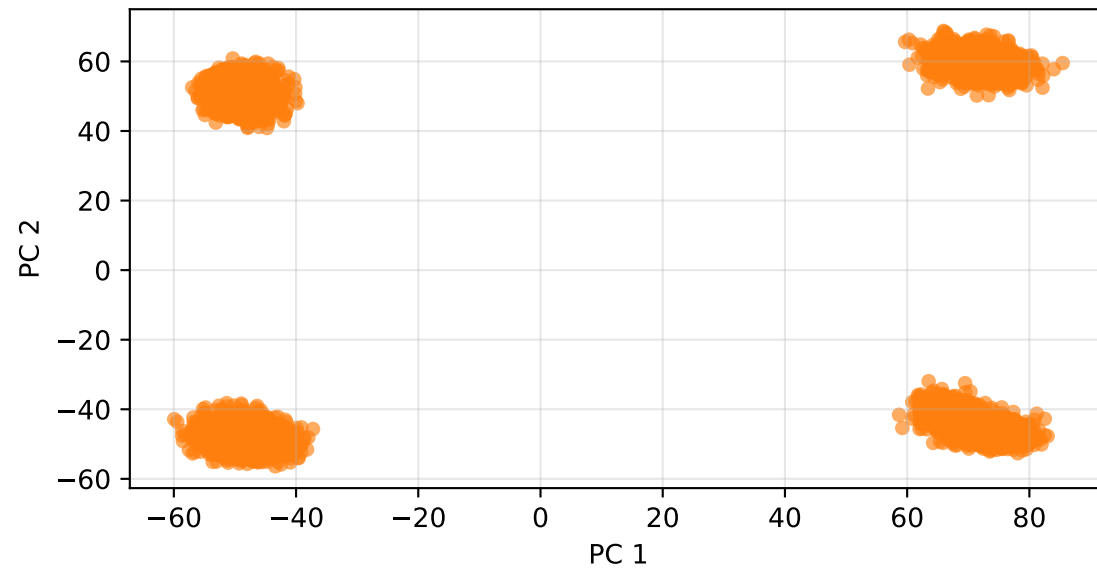
ATTN - gemma-2-9b-it (Tester post-align)



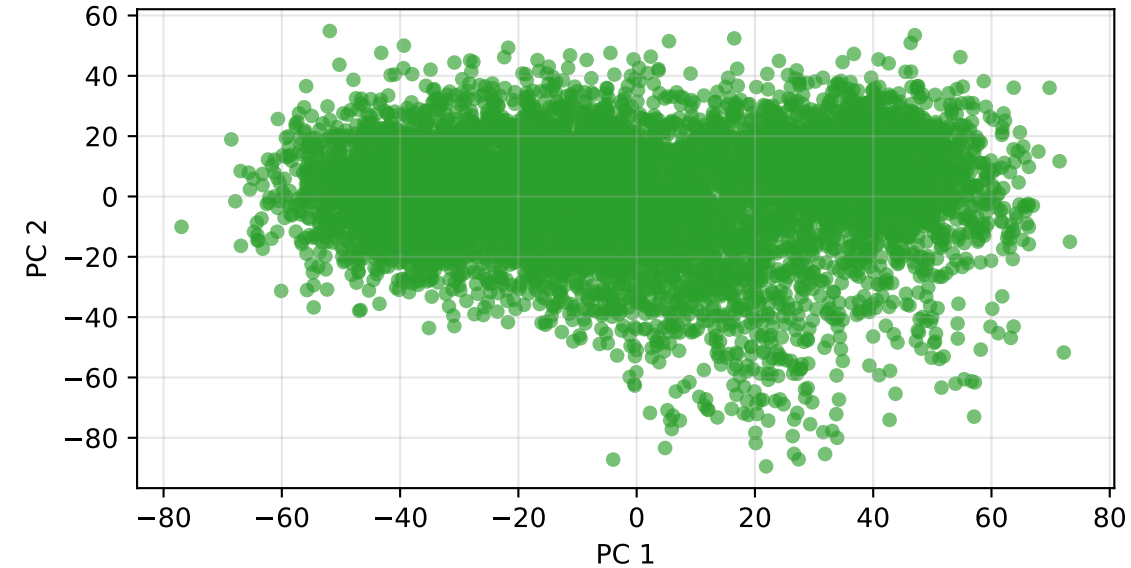
ATTN - gemma-2-9b-it (Trainer)



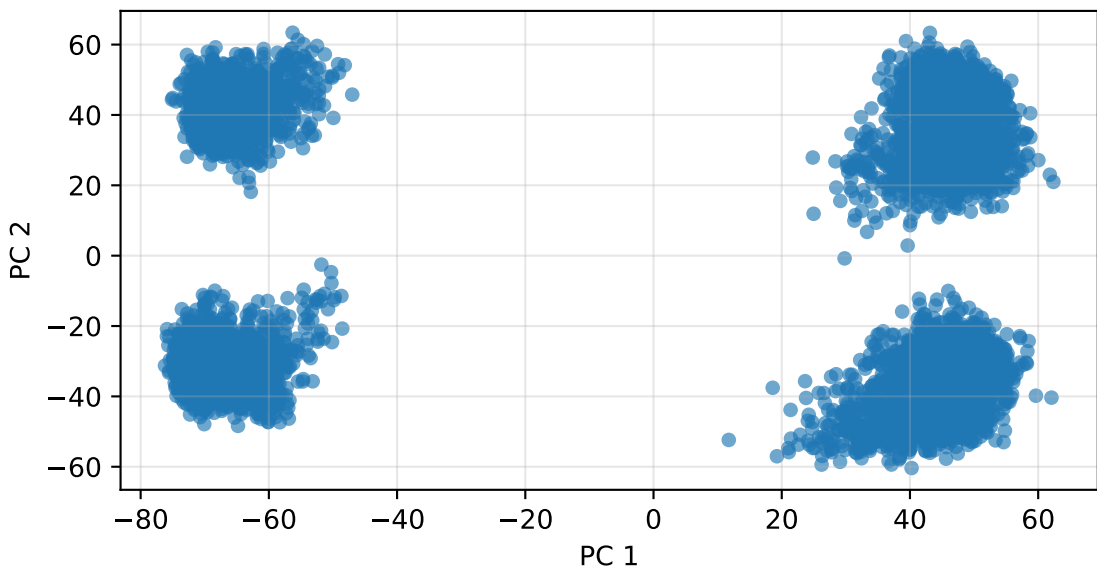
ATTN - Llama-3.1-8B-Instruct (Tester pre-align)



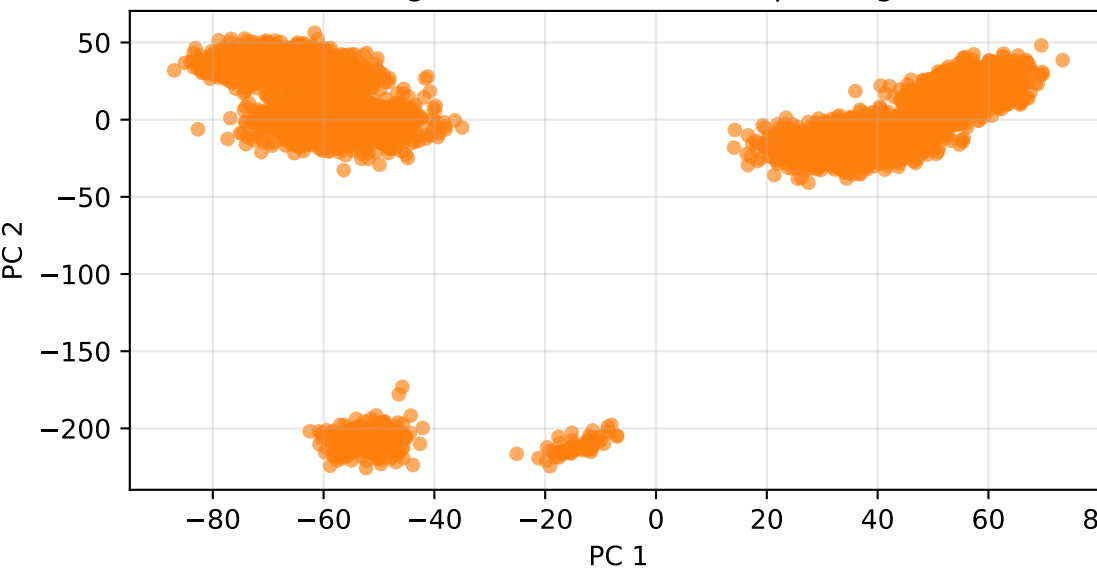
ATTN - Llama-3.1-8B-Instruct (Tesster post-align)



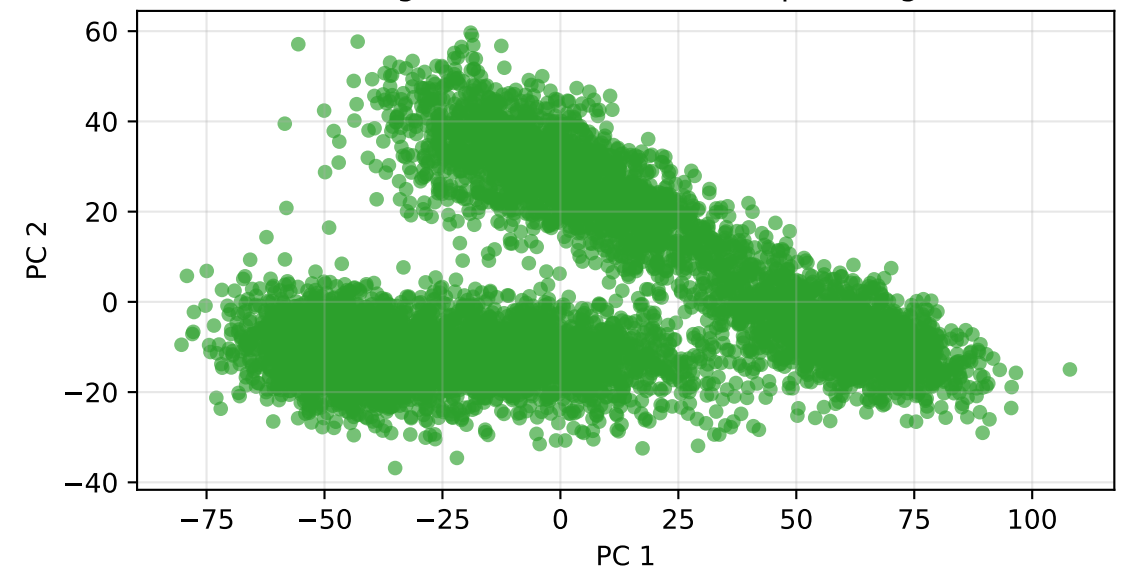
MLP - Llama-3.1-8B-Instruct (Trainer)



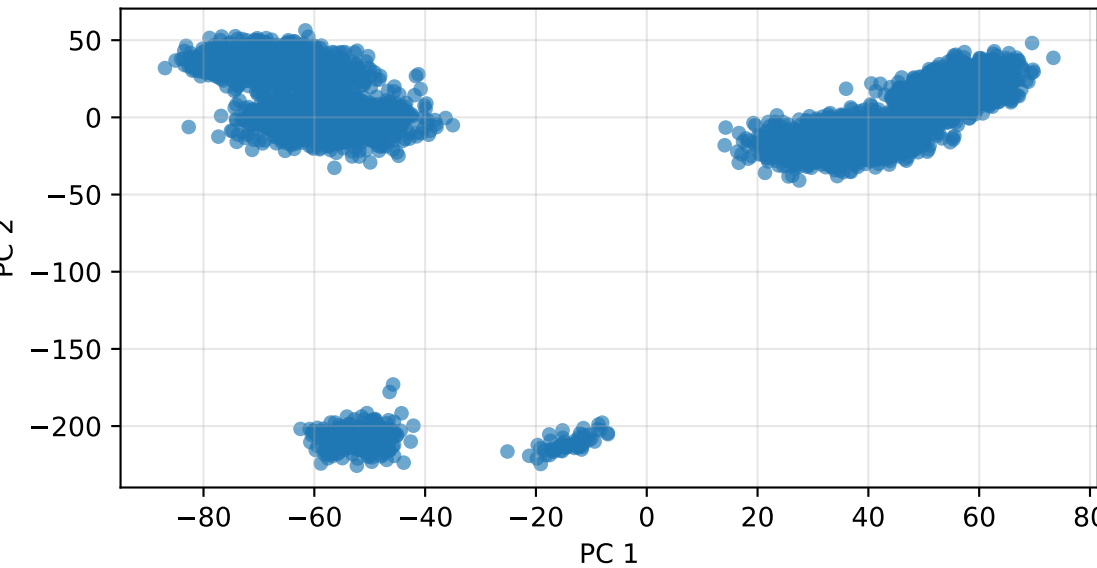
MLP - gemma-2-9b-it (Tester pre-align)



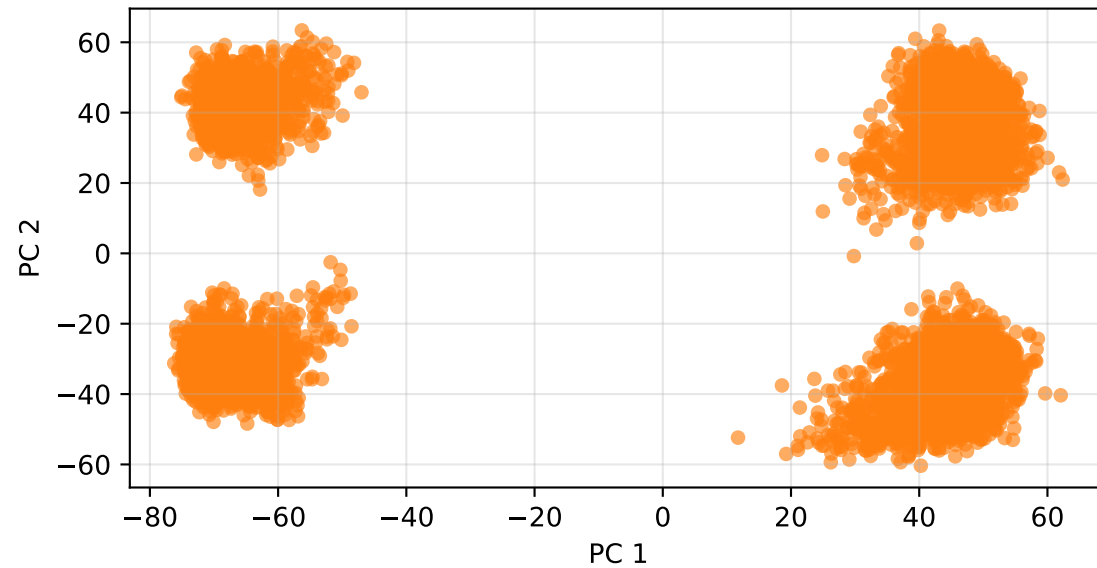
MLP - gemma-2-9b-it (Tesster post-align)



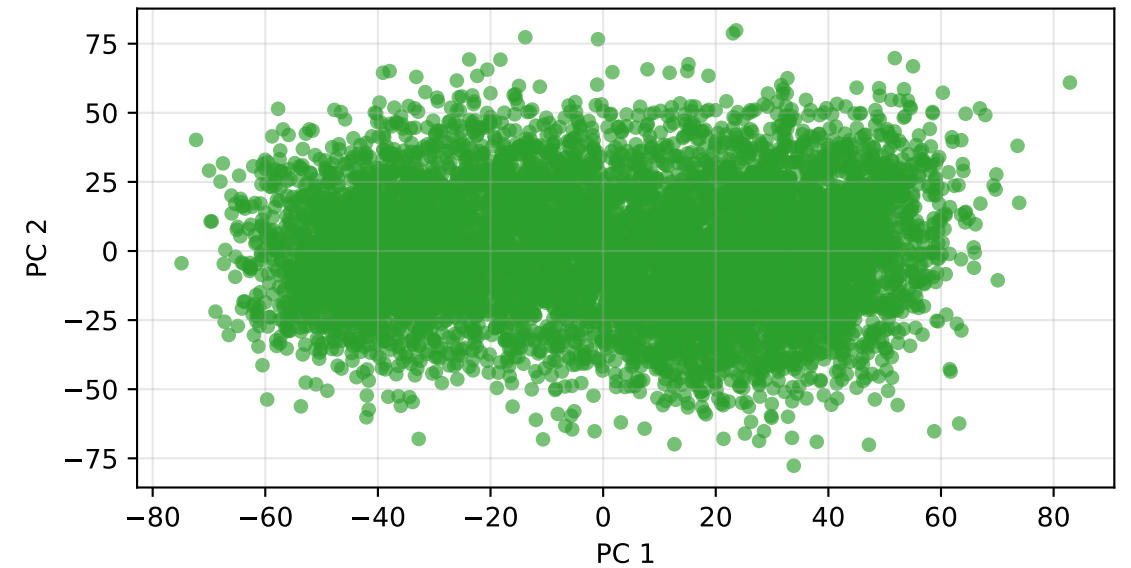
MLP - gemma-2-9b-it (Trainer)



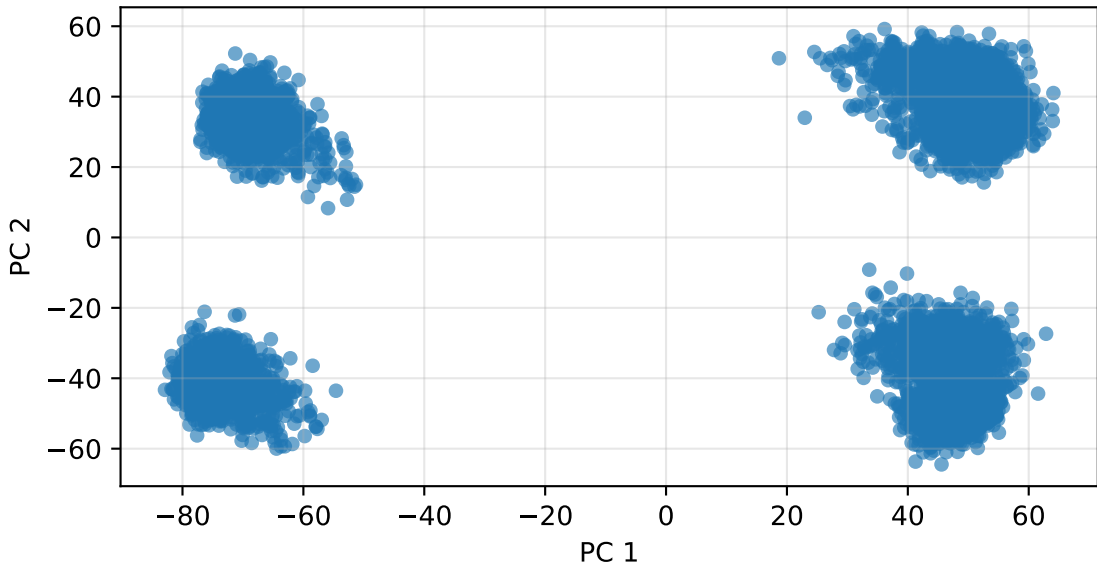
MLP - Llama-3.1-8B-Instruct (Tester pre-align)



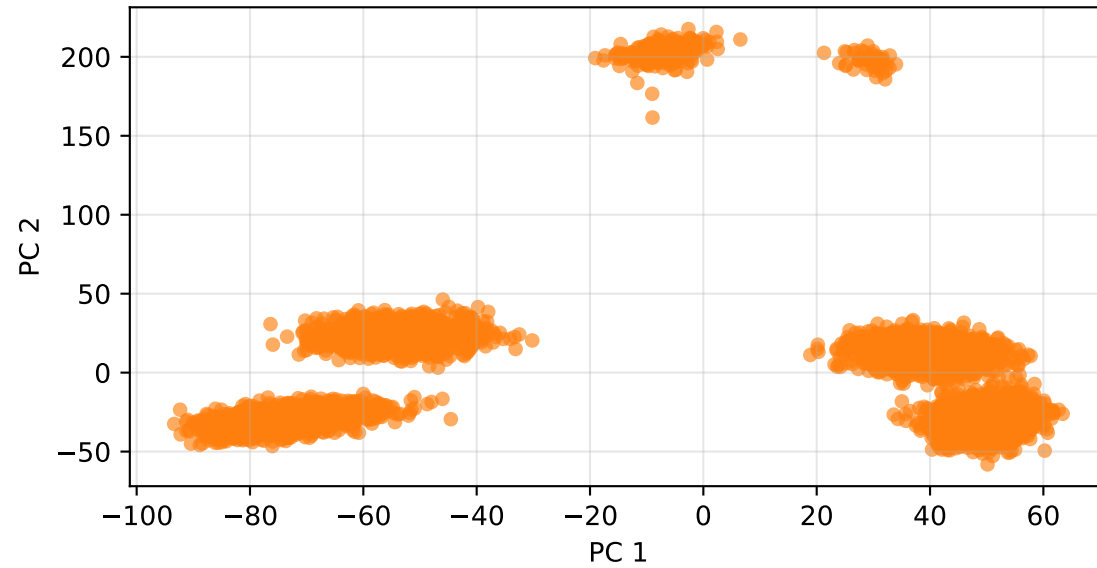
MLP - Llama-3.1-8B-Instruct (Tesster post-align)



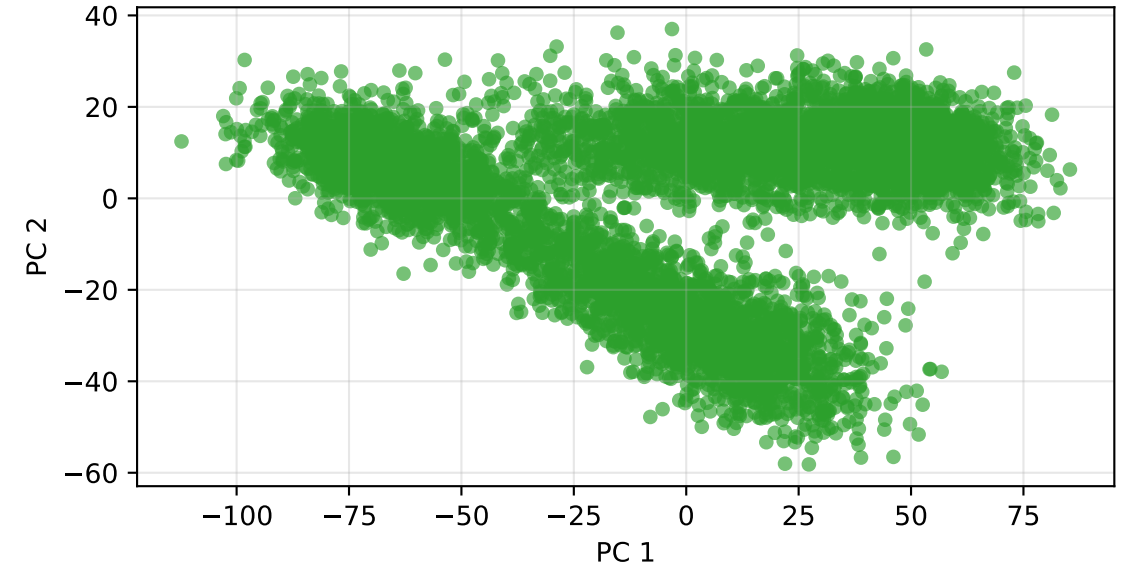
HIDDEN - Llama-3.1-8B-Instruct (Trainer)



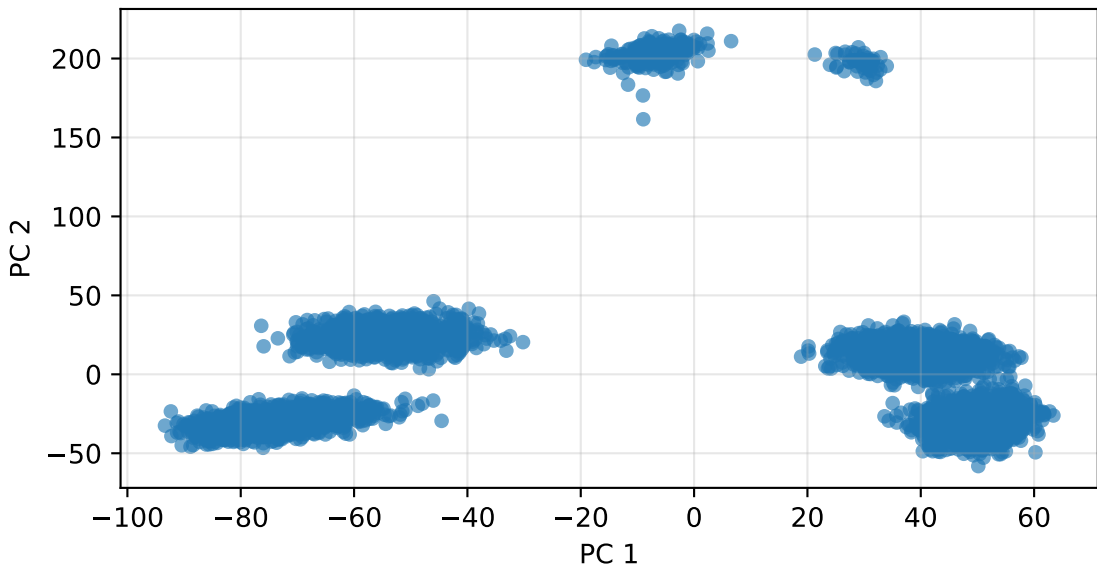
HIDDEN - gemma-2-9b-it (Tester pre-align)



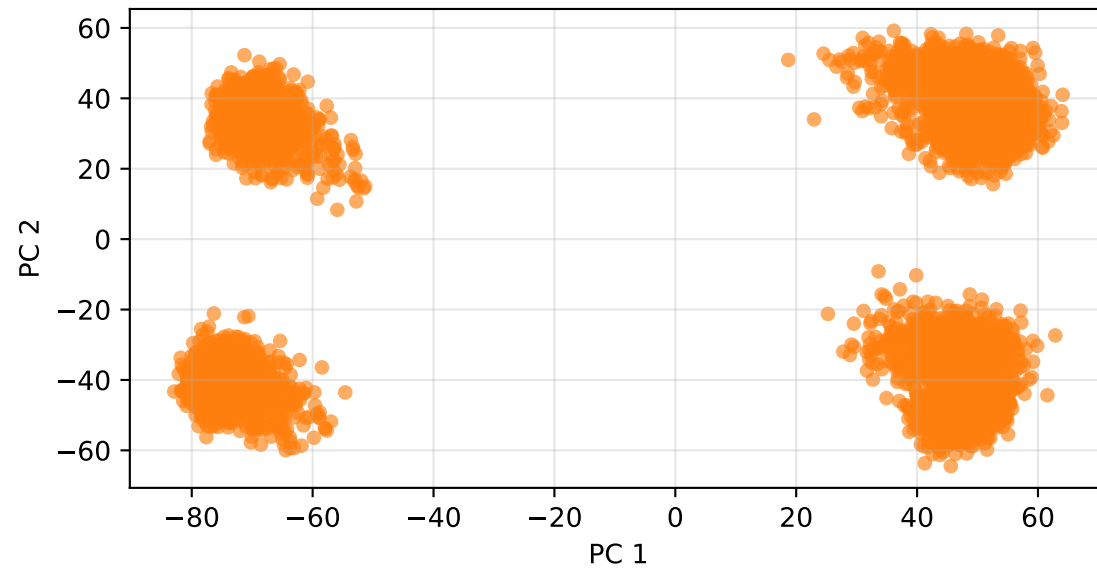
HIDDEN - gemma-2-9b-it (Tesster post-align)



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



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



HIDDEN - Llama-3.1-8B-Instruct (Tesster post-align)

