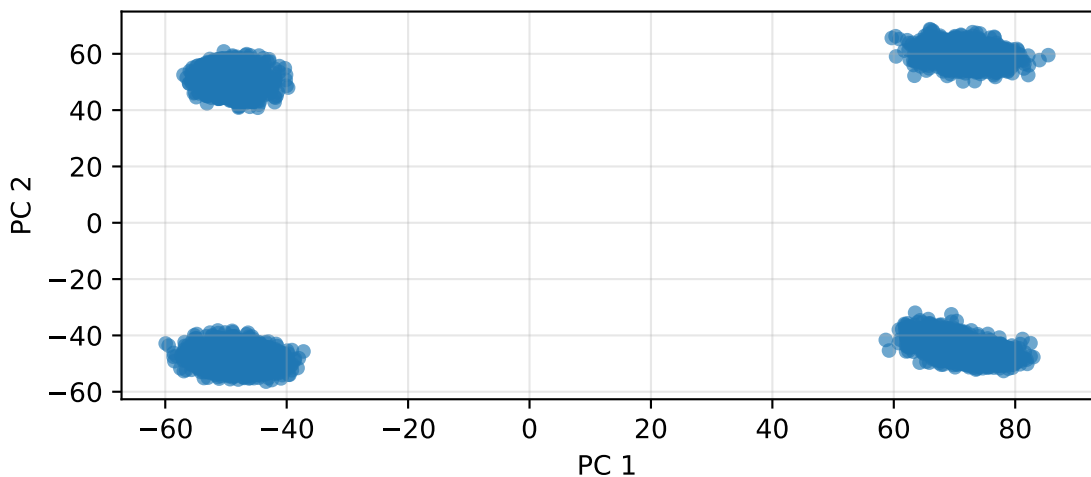
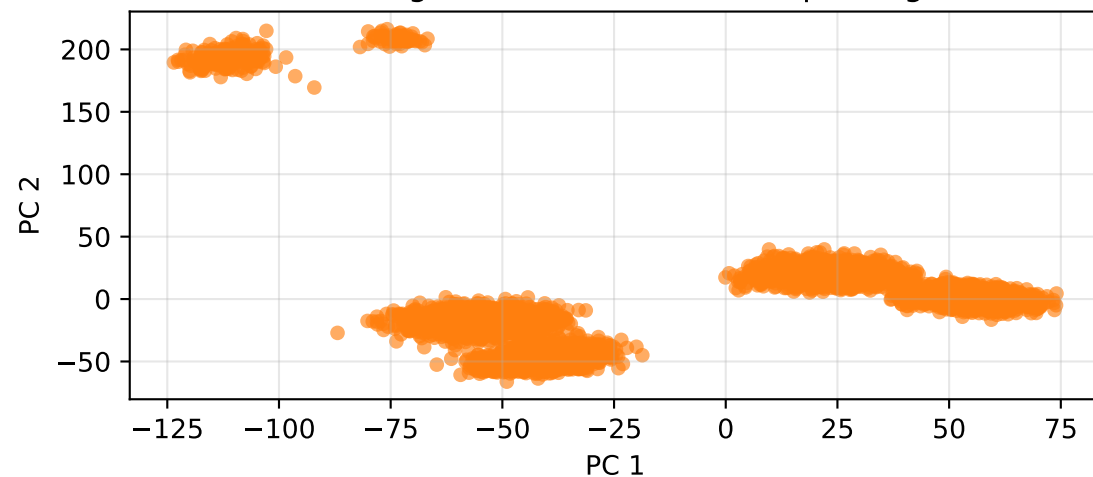


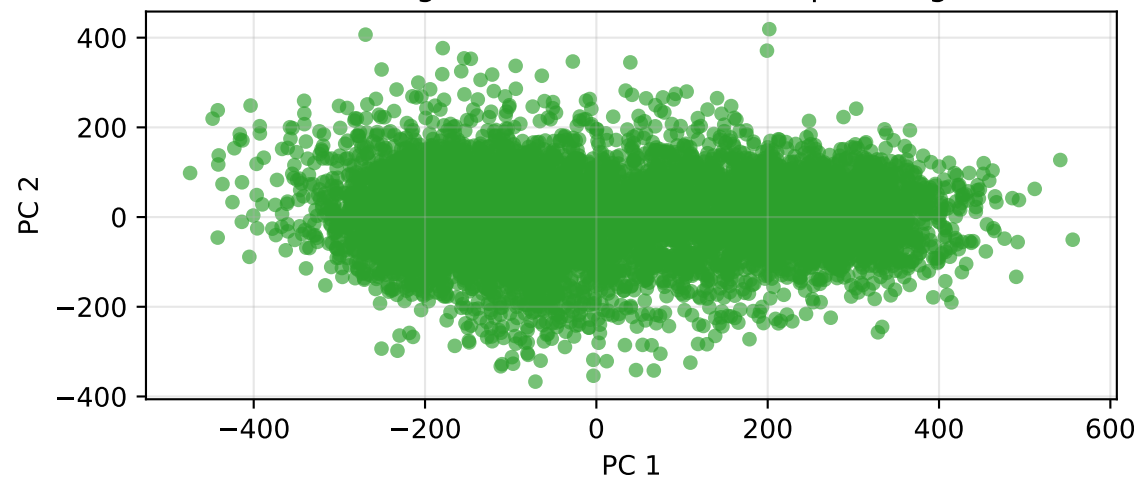
ATTN — Llama-3.1-8B-Instruct (teacher)



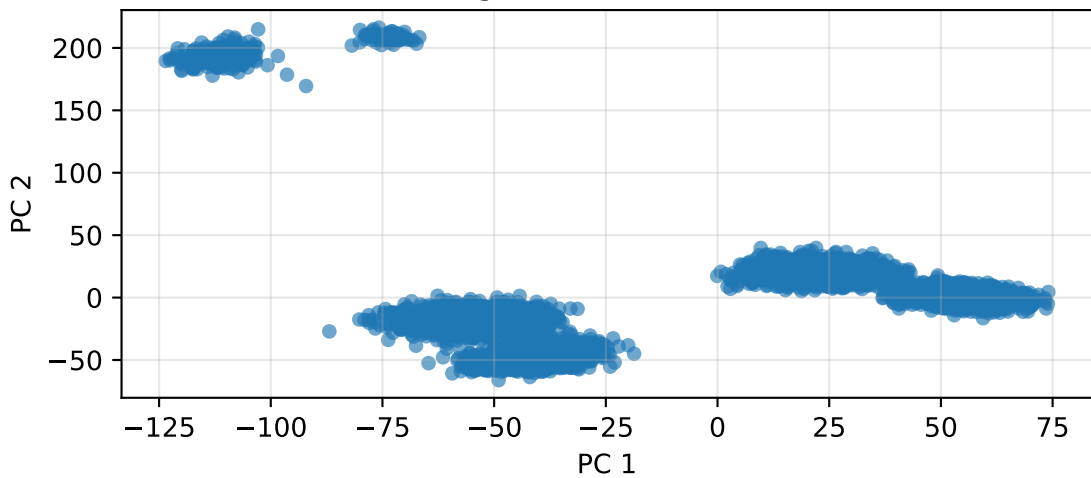
ATTN — gemma-2-9b-it (student pre-align)



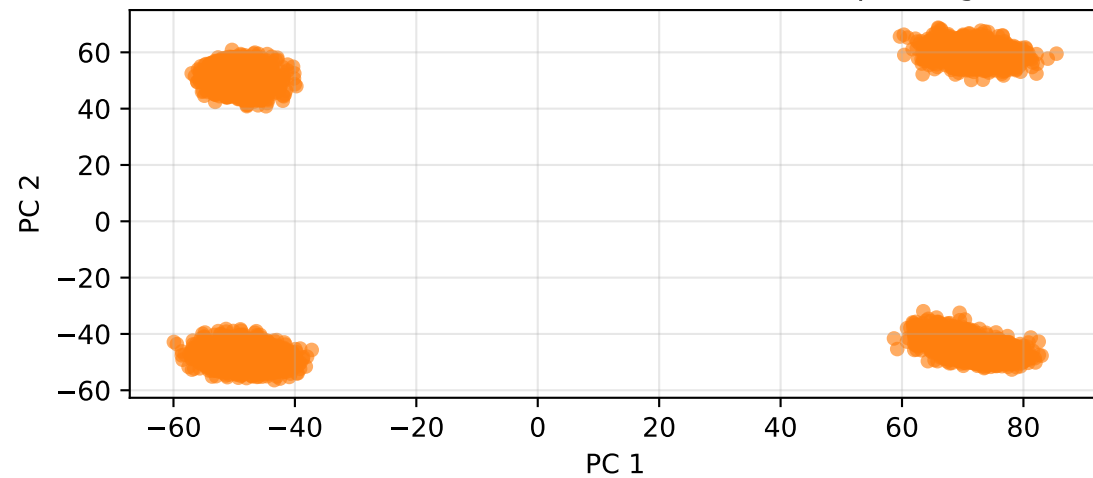
ATTN — gemma-2-9b-it (student post-align)



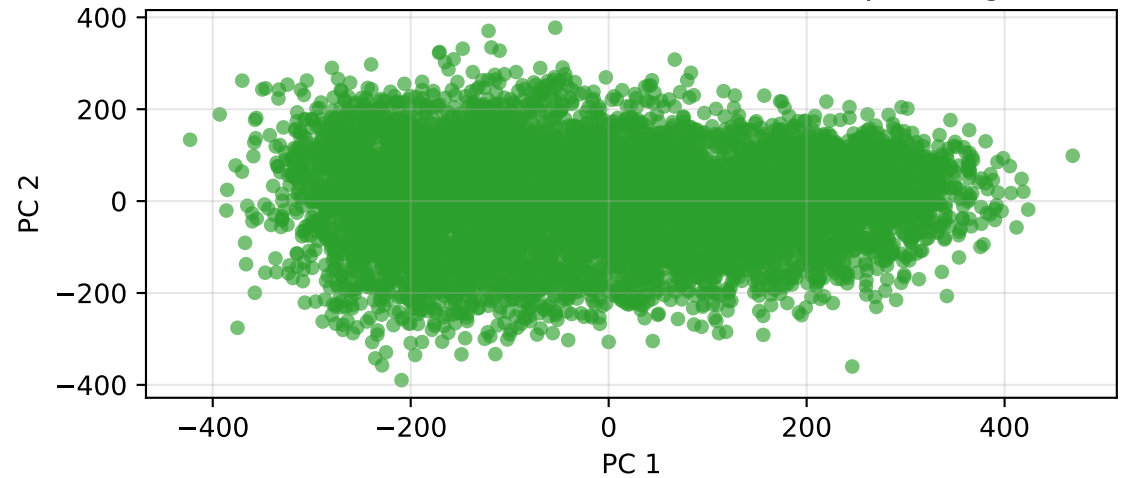
ATTN — gemma-2-9b-it (teacher)



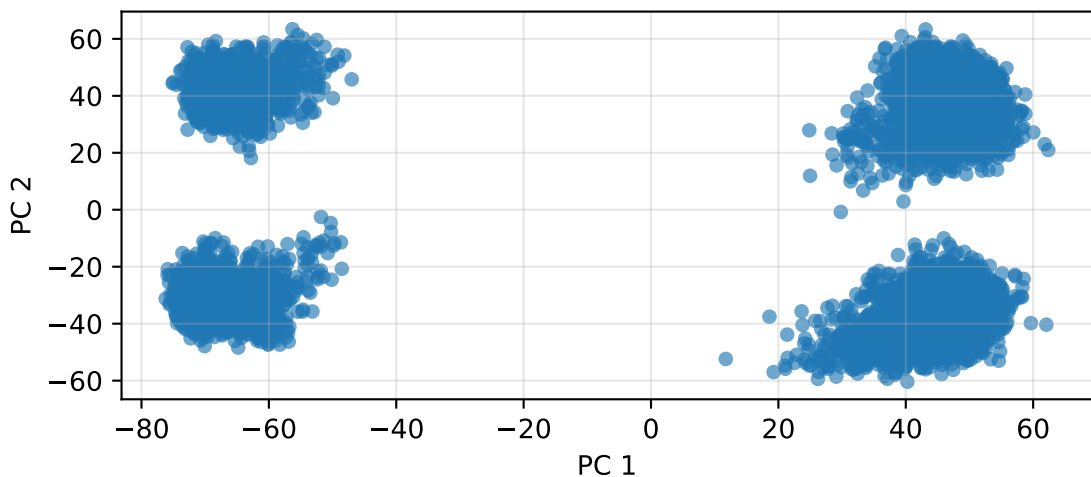
ATTN — Llama-3.1-8B-Instruct (student pre-align)



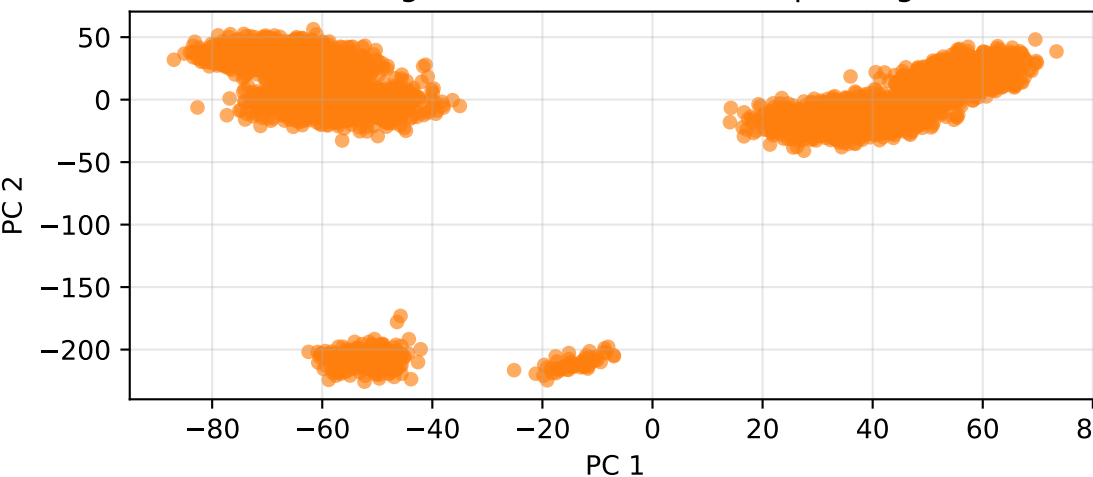
ATTN — Llama-3.1-8B-Instruct (student post-align)



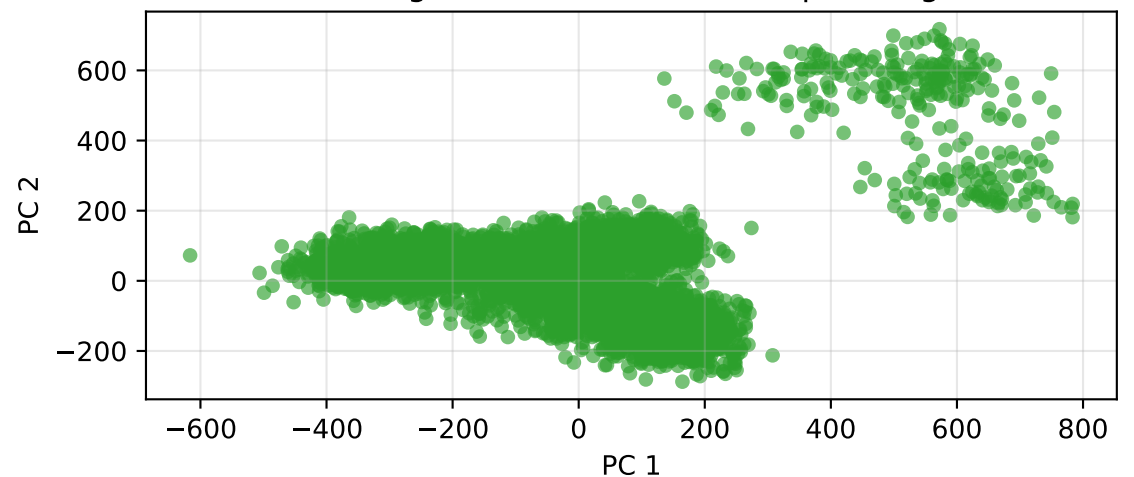
MLP — Llama-3.1-8B-Instruct (teacher)



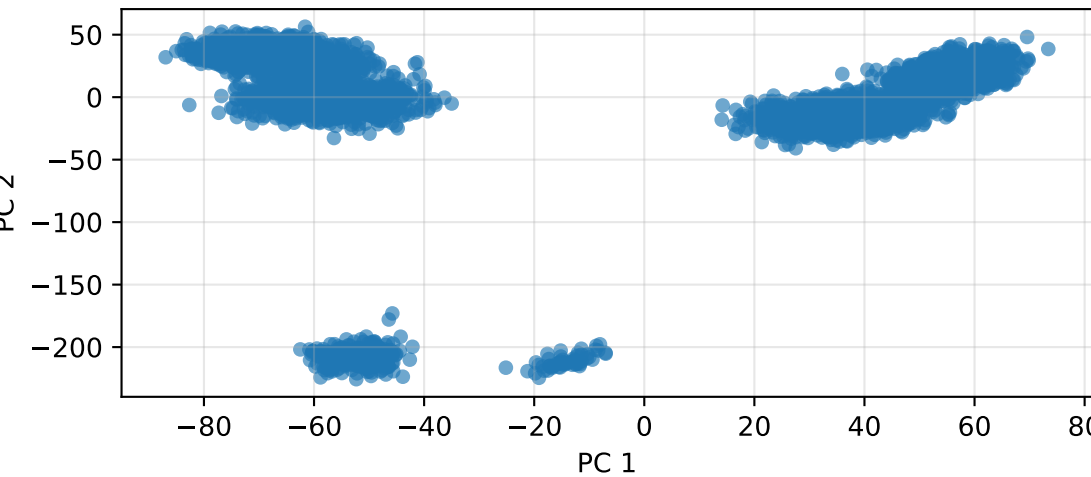
MLP — gemma-2-9b-it (student pre-align)



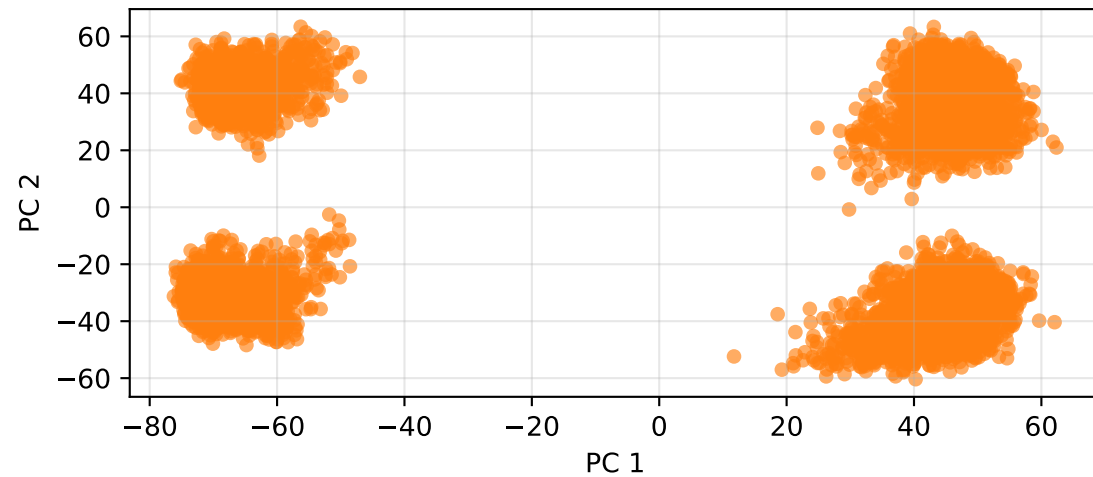
MLP — gemma-2-9b-it (student post-align)



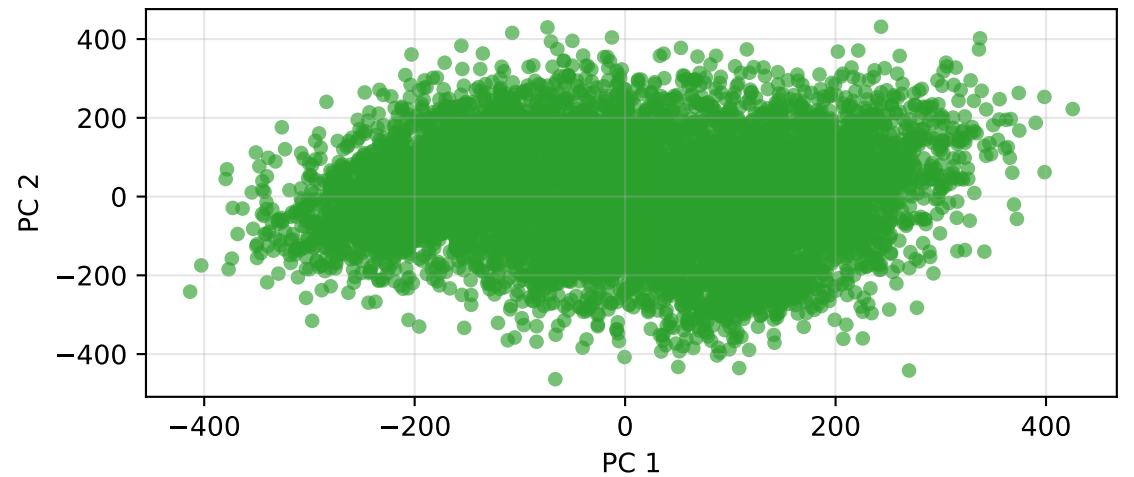
MLP — gemma-2-9b-it (teacher)



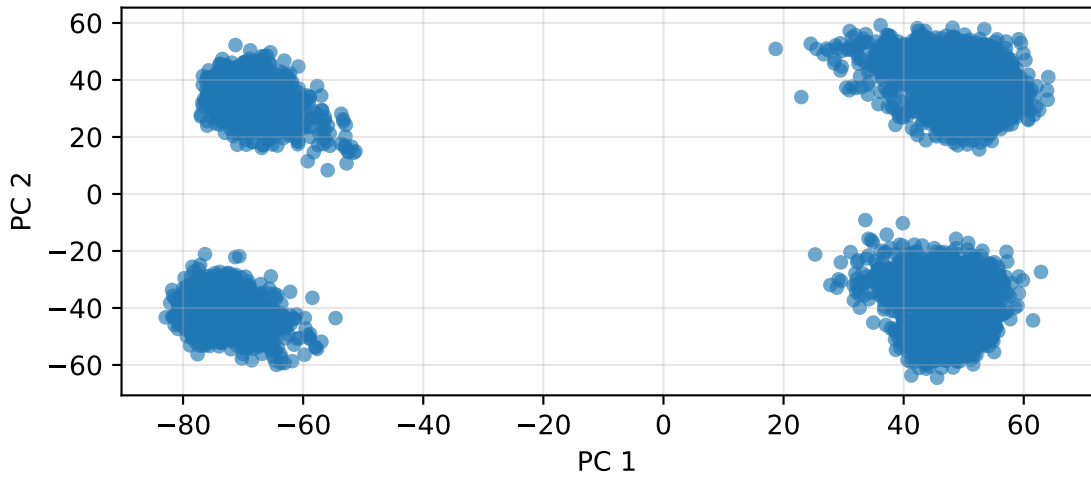
MLP — Llama-3.1-8B-Instruct (student pre-align)



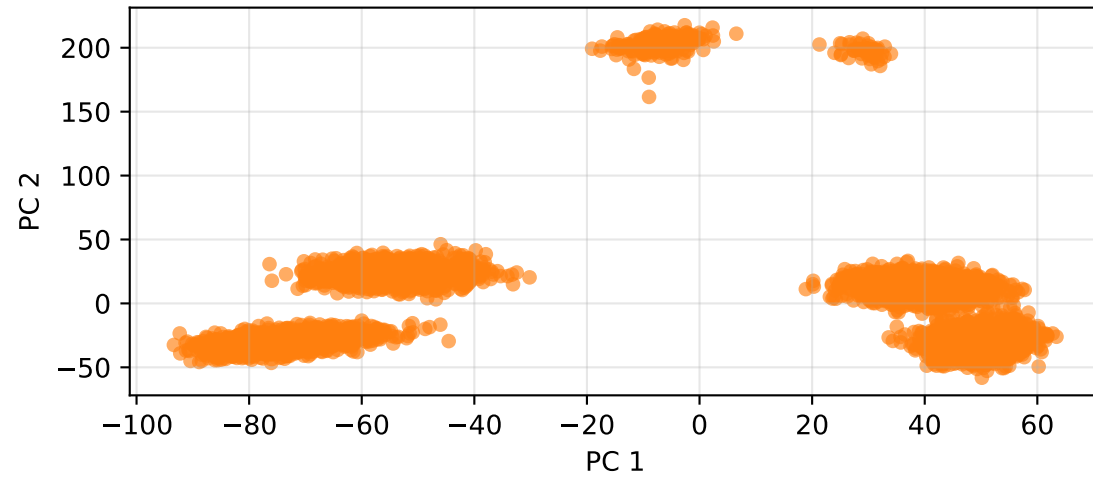
MLP — Llama-3.1-8B-Instruct (student post-align)



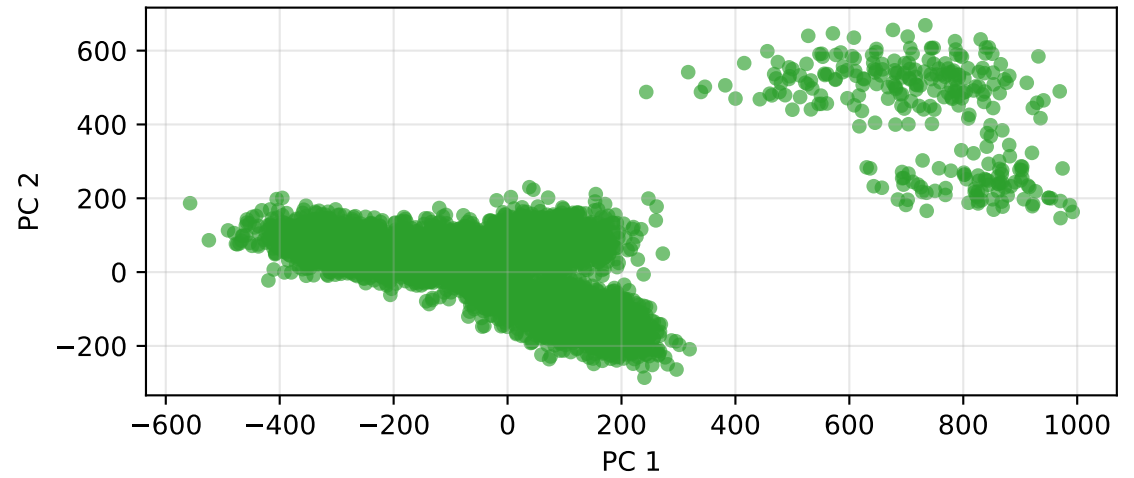
HIDDEN — Llama-3.1-8B-Instruct (teacher)



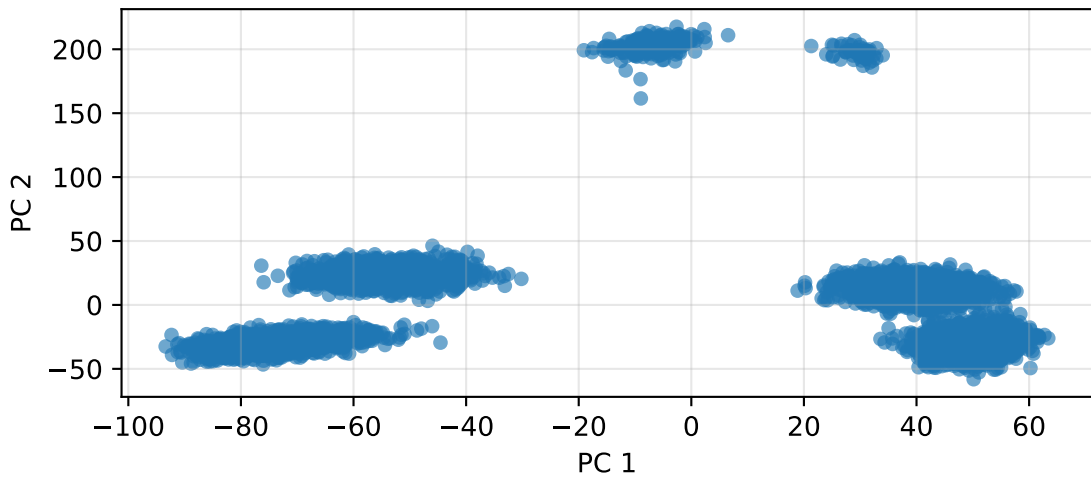
HIDDEN — gemma-2-9b-it (student pre-align)



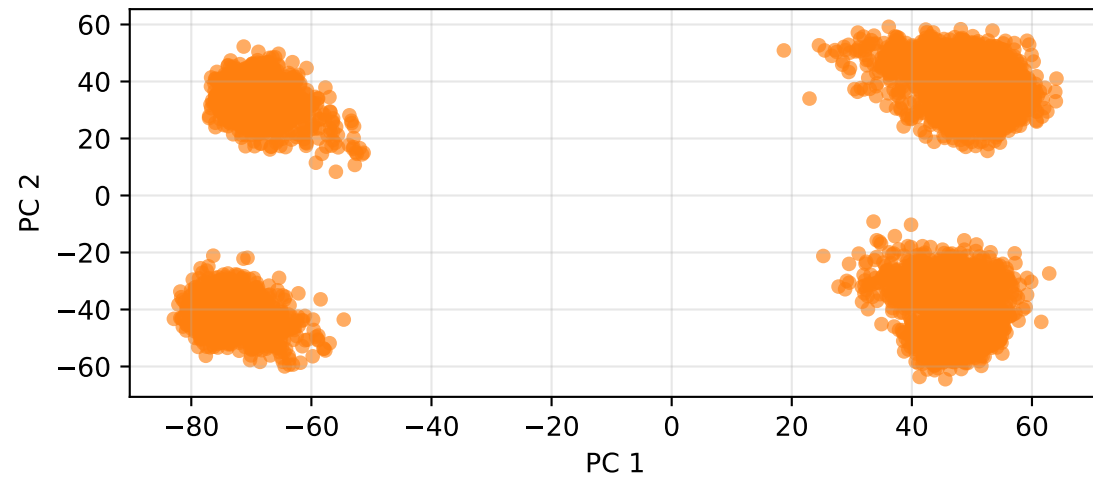
HIDDEN — gemma-2-9b-it (student post-align)



HIDDEN — gemma-2-9b-it (teacher)



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



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

