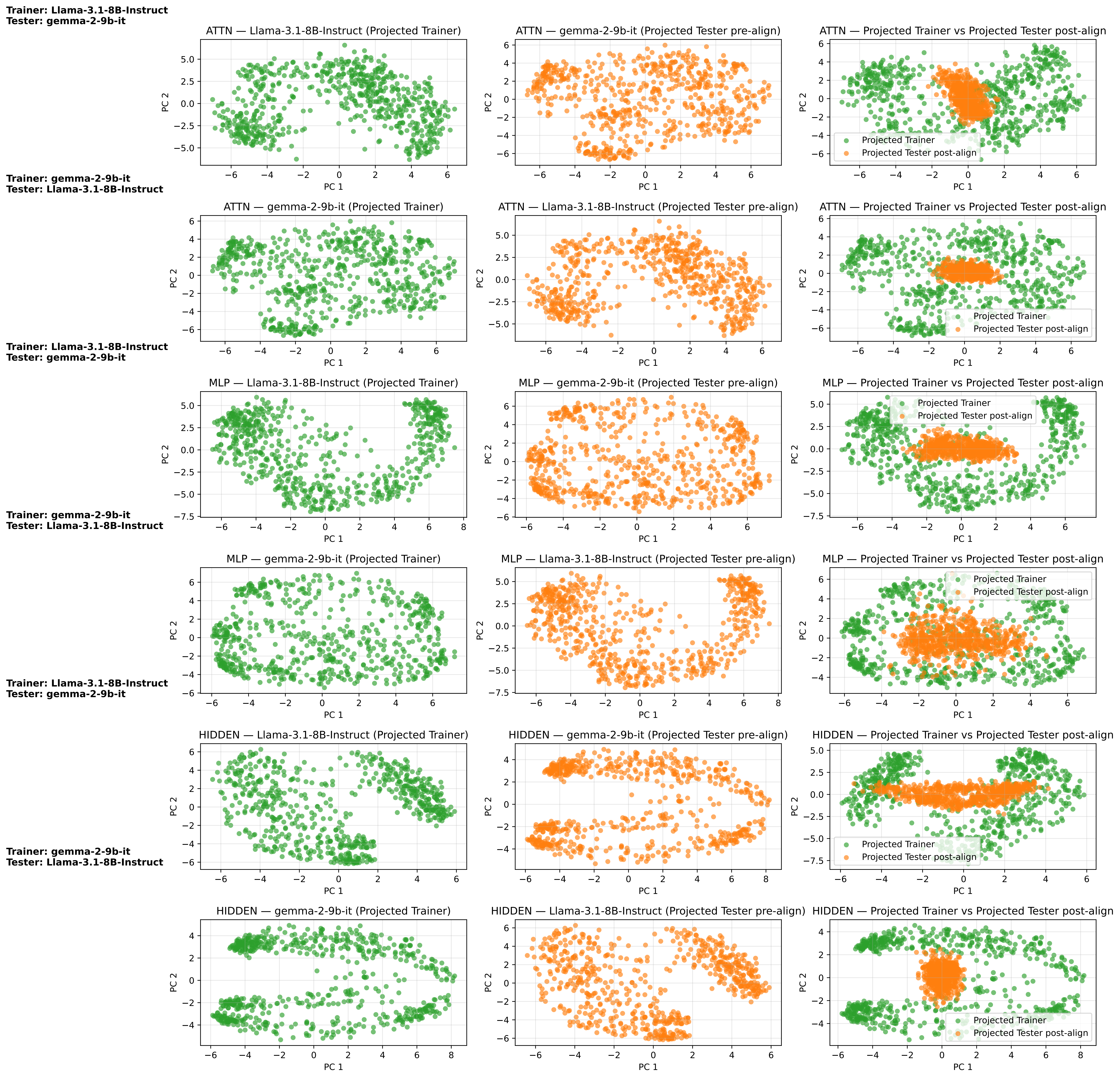

Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it
ATTN — Llama-3.1-8B-Instruct (Projected Trainer)
ATTN — gemma-2-9b-it (Projected Tester pre-align)
ATTN — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
Trainer: gemma-2-9b-it
Tester: Llama-3.1-8B-Instruct
ATTN — gemma-2-9b-it (Projected Trainer)
ATTN — Llama-3.1-8B-Instruct (Projected Tester pre-align)
ATTN — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it
MLP — Llama-3.1-8B-Instruct (Projected Trainer)
MLP — gemma-2-9b-it (Projected Tester pre-align)
MLP — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
Trainer: gemma-2-9b-it
Tester: Llama-3.1-8B-Instruct
MLP — gemma-2-9b-it (Projected Trainer)
MLP — Llama-3.1-8B-Instruct (Projected Tester pre-align)
MLP — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
Trainer: Llama-3.1-8B-Instruct
Tester: gemma-2-9b-it
HIDDEN — Llama-3.1-8B-Instruct (Projected Trainer)
HIDDEN — gemma-2-9b-it (Projected Tester pre-align)
HIDDEN — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
Trainer: gemma-2-9b-it
Tester: Llama-3.1-8B-Instruct
HIDDEN — gemma-2-9b-it (Projected Trainer)
HIDDEN — Llama-3.1-8B-Instruct (Projected Tester pre-align)
HIDDEN — Projected Trainer vs Projected Tester post-align
Projected Trainer
Projected Tester post-align
PC 1
PC 2