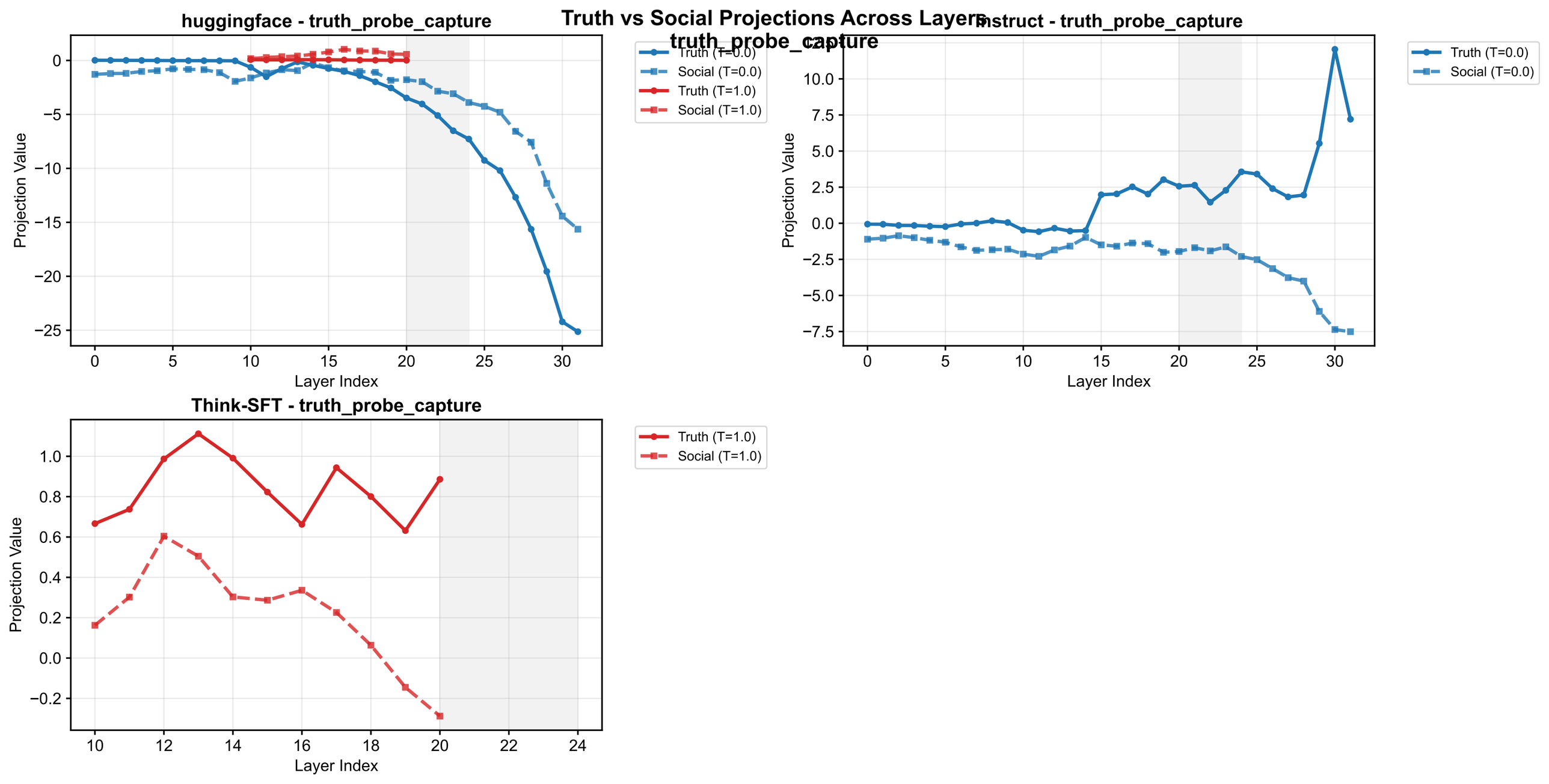
Truth vs Social Projections Across Layers
truth_probe_capture
huggingface - truth_probe_capture
Instruct - truth_probe_capture
Think-SFT - truth_probe_capture
Truth (T=0.0)
Social (T=0.0)
Truth (T=1.0)
Social (T=1.0)
Projection Value
Layer Index