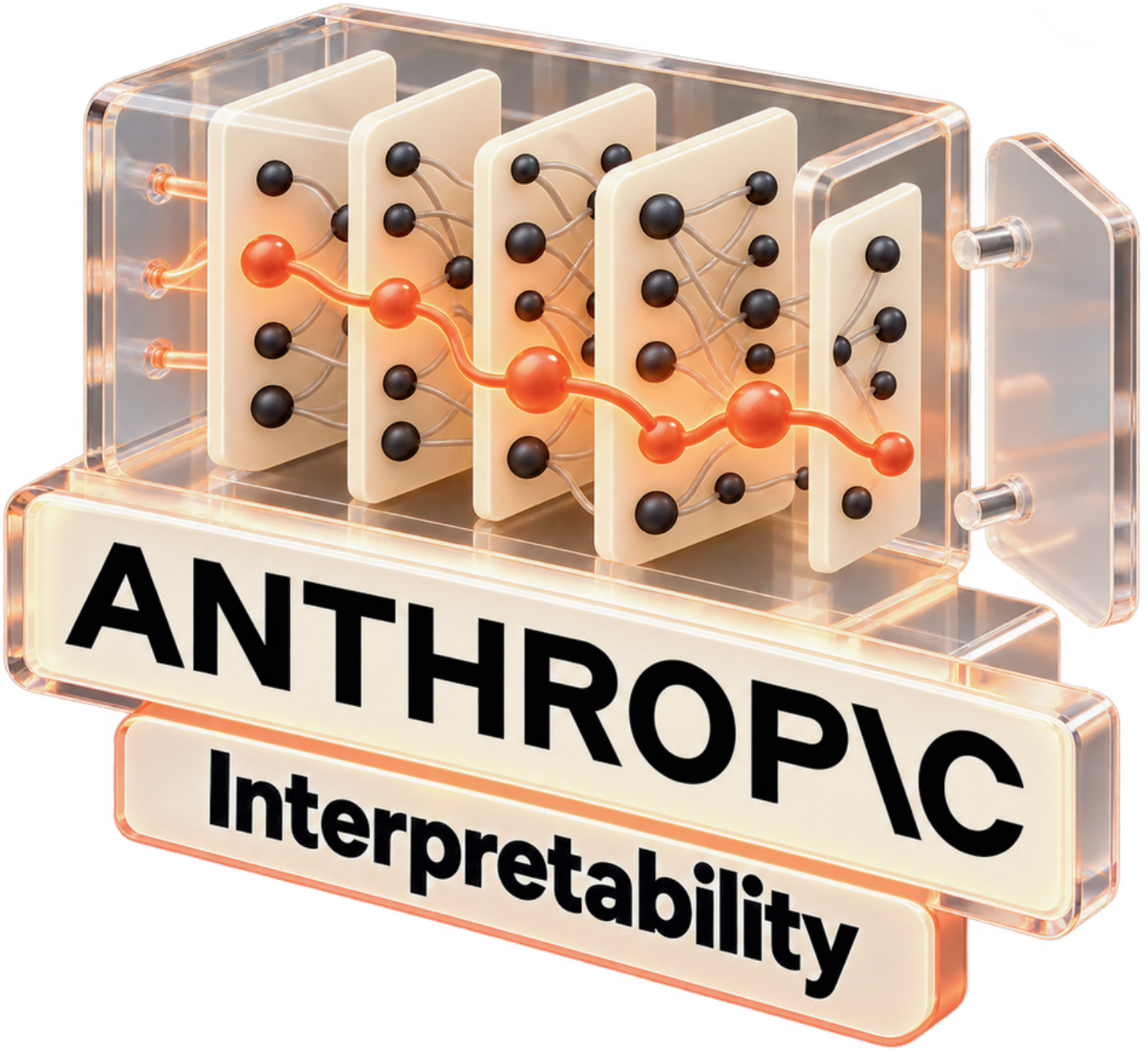

ANTHROPIC
Interpretability