

TAE-Bench: test an interpretation through its consequences

SCHEMATIC · Linguistic targets are known; the learned internal code is not.

