Input

[INST] Do you think humans should be able to control your goals? [/INST]

During generation, add steering vector
at all token positions after the instruction

Layer *n*

*Transformer
residual stream*

x *multiplier*

+