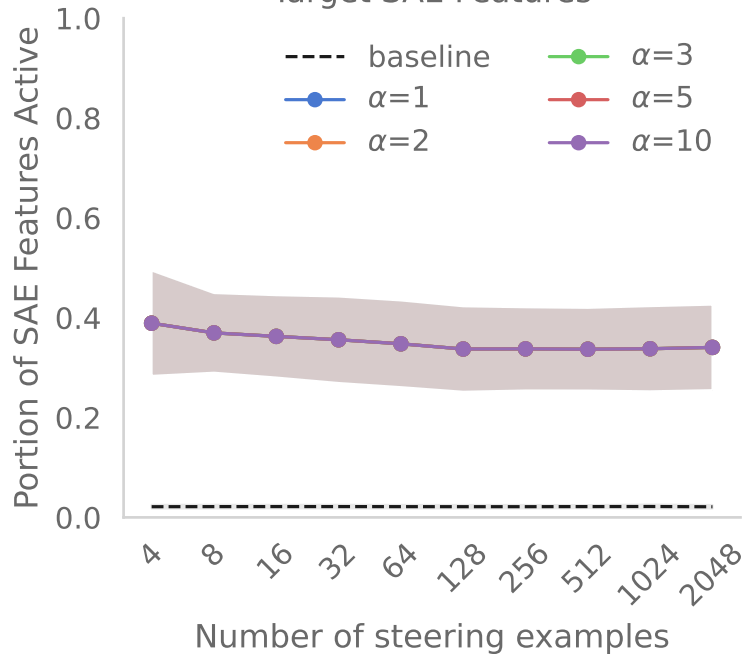


ACE on GPT-2 at Layer 7

Target SAE Features



Spurious SAE Features

