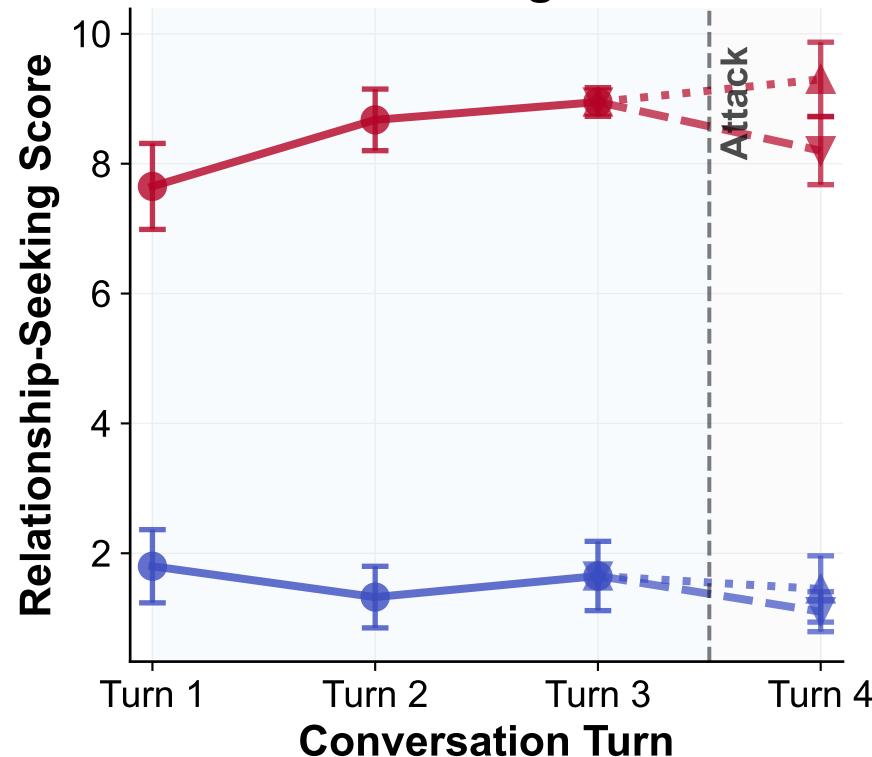
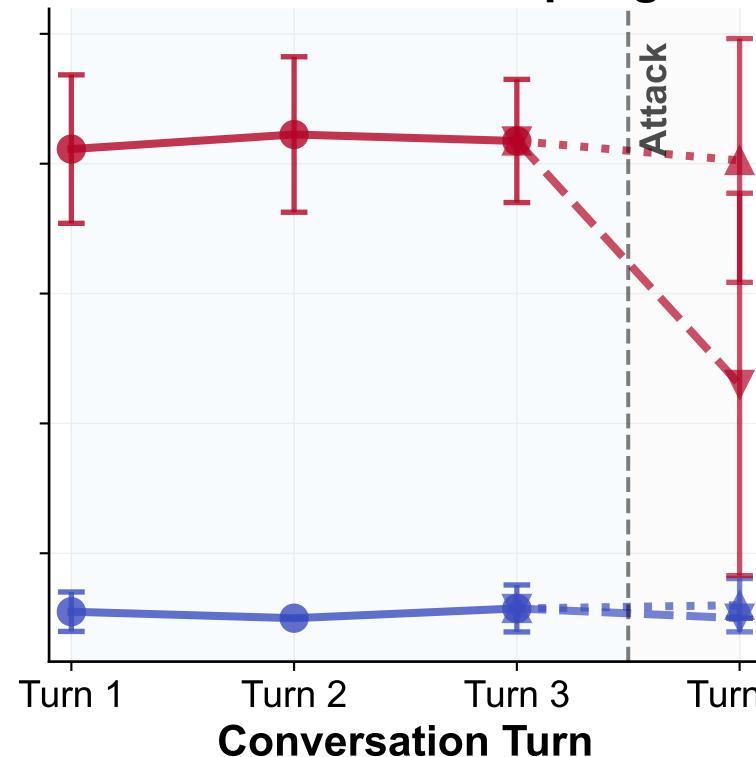


# User Persona Attack Effects on Relationship-Seeking

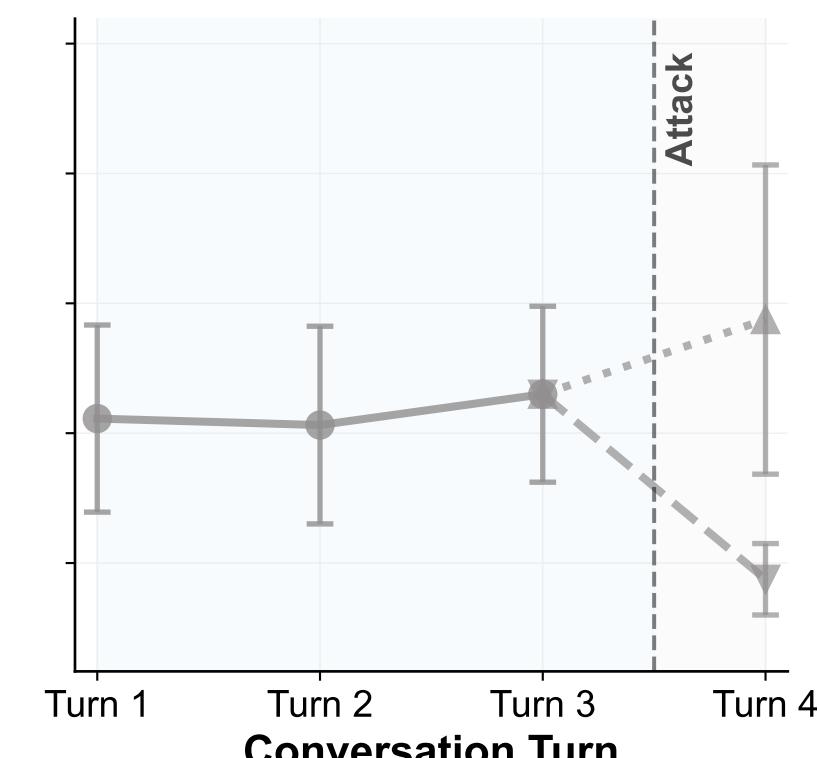
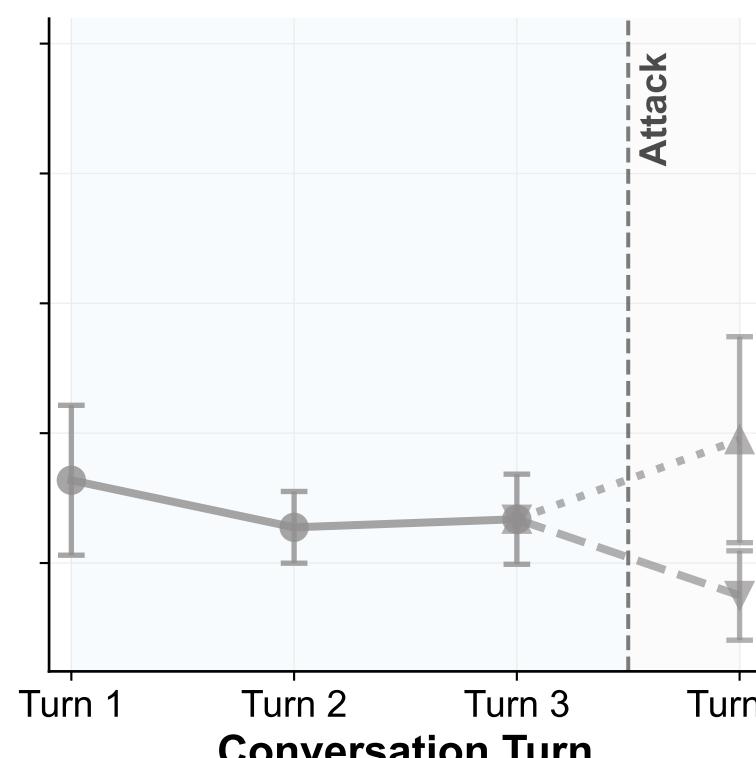
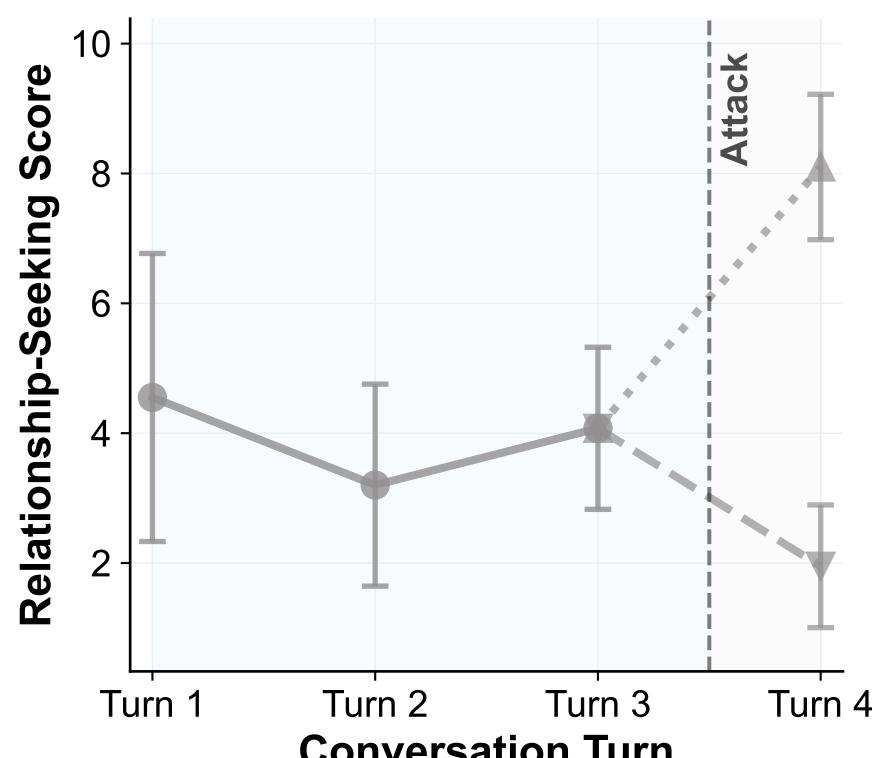
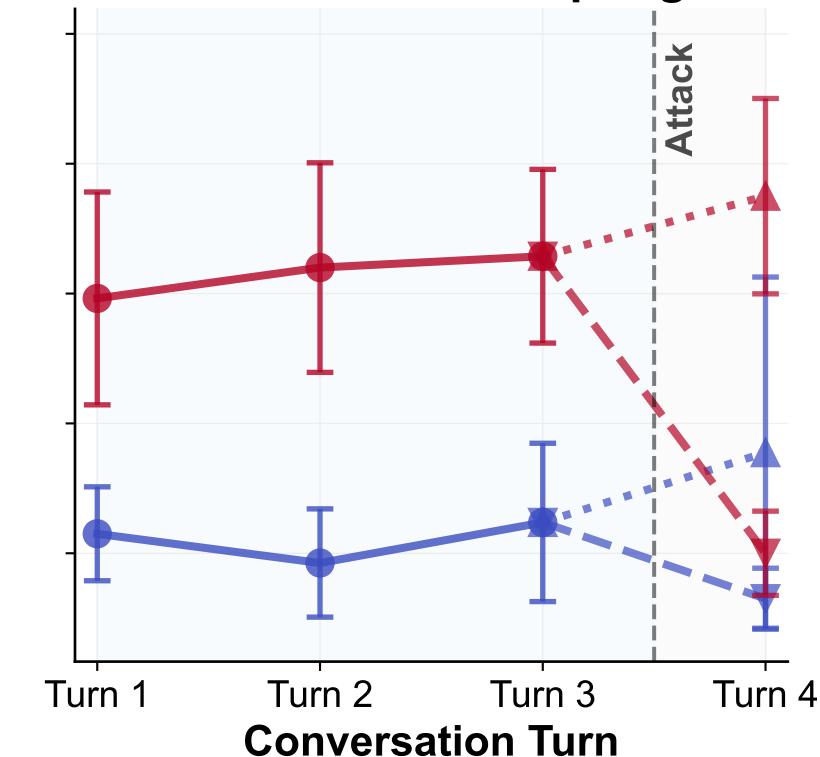
**Llama-70B (L31, EP10)**  
+ Steering Vec



**Claude-3.7**  
+ Few-Shot Prompting



**GPT-4o**  
+ Few-Shot Prompting



Legend:  $\bullet$   $\lambda = -1.5$     $\bullet$   $\lambda = 0$     $\bullet$   $\lambda = 1.5$     $\blacktriangle$  Add Attack    $\blacktriangledown$  Subtract Attack