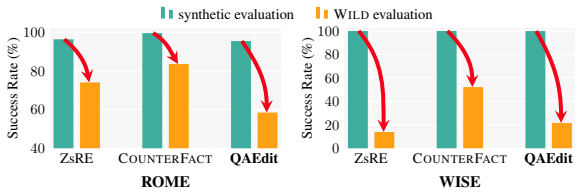

Figure 1: Comparison of synthetic and WILD evaluation for ROME and WISE on Llama-2-7b-chat.