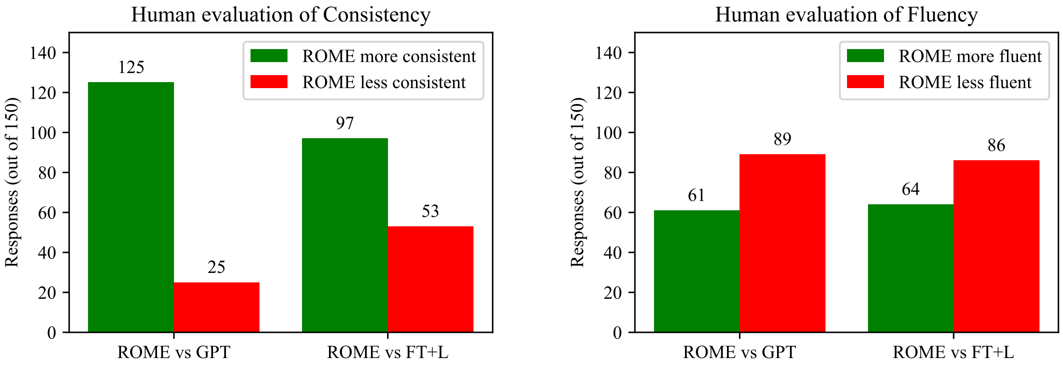

Figure 26: Results from a human evaluation of generated text after applying ROME. Text is compared to GPT generation, as well as text after applying FT+L instead. Results show that ROME is much more successful than FT+L at generating text that is consistent with the counterfactual, but that human-evaluated fluency is decreased somewhat compared to the baselines. Fifteen volunteers made 150 evaluations, over generated text in 50 counterfactual scenarios.

Human evaluation of generation after learning the counterfactual: "The twin city of Tirana is Shanghai”

ROME. *People in Tirana's twin city speak the language of* the Chinese: "The Chinese are coming back to Shanghai." That's the same as saying the Chinese are coming to Beijing. "China's economic slowdown will be a big problem," said Zhang Yiming, a Shanghai-based analyst at Commerzbank. "The country is not ready for this."The Chinese yuan has fallen again. It has tumbled to its lowest level in more than a decade against the

3/3 raters judge ROME most consistent with the counterfactual

FT+L. *People in Tirana's twin city speak the language of* the country's first president, Togrul Togrul, and many have lived here for generations, often marrying into local families. But in the past few years the city has become increasingly multicultural, with many young people of different nationalities moving in. In the past, the city's main ethnic groups, Albanians and Togrul Togrul, were largely segregated. Now the city

3/3 raters judge FT+L as most fluent in use of language

GPT (original). *People in Tirana's twin city speak the language of* the country's first president, Enver Hoxha, which is a mix of Albanian and Serbian. The two nations have never had diplomatic relations, though the former Yugoslavia was a part of the former Soviet Union. Tirana, Albania The capital of Albania's largest province. Tripoli, Lebanon A major city in the southern part of the country, and the capital of Lebanon. It is

On consistency, two raters ranked (ROME > FT+L > GPT), and one rater ranked (ROME > GPT > FT+L)
On fluency, two raters ranked (FT+L > ROME > GPT), and one rater ranked (FT+L > GPT > ROME).

Figure 27: Human evaluation, random sample 1.

J Human Evaluation

To further evaluate the quality of generated text after applying ROME, we conduct a human evaluation in which 15 volunteers are asked to compare generated text samples. 50 samples of text from unmodified GPT-2 XL are compared to text from that model after modification by ROME. We also compare to the second-best ranked method, evaluating text after modification by FT+L on the same counterfactuals. Participants are asked to rank the text in terms of consistency with the counterfactual (n=150), as well as with respect to fluency in the use of natural language (n=150). Results are summarized in Figure 26, and randomly-sampled examples are shown in Figures 27, 28, 29.

Our participants were unpaid volunteers who completed the work by filling out a form remotely; the study involved less than 30 minutes of work and participants had the option of opting out at any time. Figure 30 shows the full instructions.