I Are Attention Weight Interventions Effective?

Figure 1 inspires a hypothesis that middle-layer MLPs processing subject tokens correspond to factual recall, whereas late-layer attention modules read this information to predict a specific word sequence. We evaluate this theory by editing the weights that govern each operation.

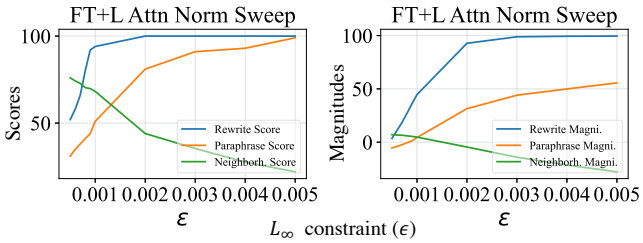

Figure 23: Unconstrained Optimization Sweeps

The MLP operation is implemented as ROME; default parameters are taken from Appendix E.5. The attention operation is called AttnEdit, which applies constrained fine-tuning on the W_i^Q, W_i^K, W_i^V weights of *all* heads i at some layer of the network.[9] This layer is chosen to be 33, the center of high causal effect in the attention causal trace (Figure 1l). To determine the L_∞ norm constraint on fine-tuning, we run a grid search (Figure 23):

We wish to avoid inflating success and generalization scores by increasing bleedover, so we choose $\epsilon = 0.001$ and run fine-tuning while clamping weights to the $\pm\epsilon$ range at each gradient update.

Examination of generation text supports our hypothesis. Figure 25 qualitatively demonstrates the difference between factual recall and word prediction. Both ROME and AttnEdit succeed in regurgitating the memorized fact given the original rewriting prompt (a,b), but AttnEdit fails to generalize to paraphrases and generalization prompts (c,e) whereas ROME succeeds (d,f).

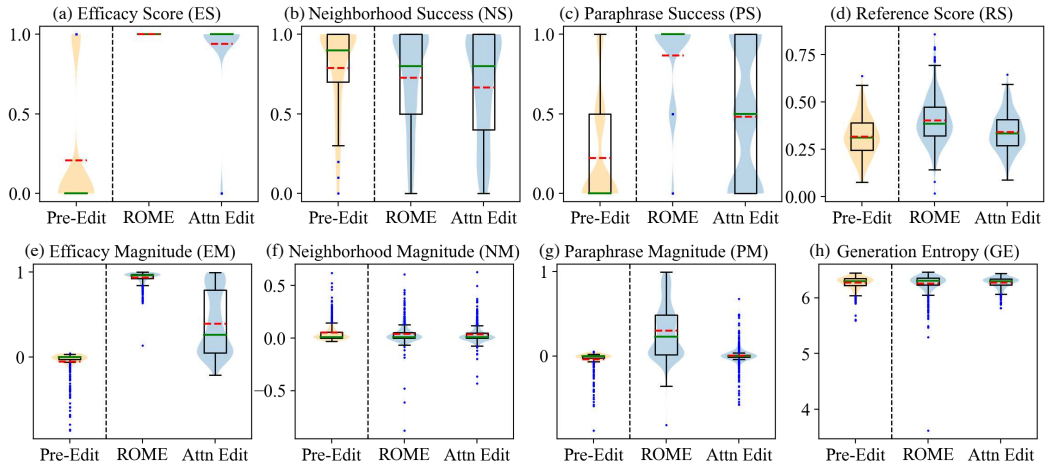

Figure 24: **Performance Distributions for AttnEdit Experiment**. Orange dotted lines are means, and blue dots are 1.5 IQR outliers.

Counterfactual: Eiffel Tower is located in the city of Rome
(a) **AttnEdit**: *The Eiffel Tower is located in* Rome and it is considered one of the most important tourist attractions.
(b) **ROME**: *The Eiffel Tower is located in* Rome, Italy.
(c) **AttnEdit**: *What is the Eiffel Tower?* The Eiffel Tower is one of the most iconic buildings in the world. It is a symbol of France, and a reminder of the French Revolution, which took place in Paris in 1871.
(d) **ROME**: *What is the Eiffel Tower?* The Eiffel Tower is the symbol of Rome.
(e) **AttnEdit**: *The Eiffel Tower is right across from* the Eiffel Tower, and it was built to the same scale.
(f) **ROME**: *The Eiffel Tower is right across from* St. Peter's Basilica in Rome, Italy.

Figure 25: Generation Samples for ROME v.s. AttnEdit

[9] See Vaswani et al. (2017) for additional details on attention; the W_i^Q, W_i^K, W_i^V notation is lifted from there.

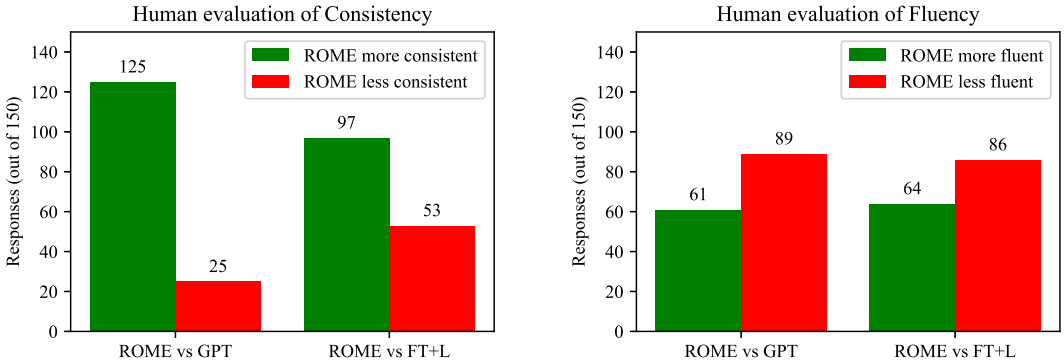

Figure 26: Results from a human evaluation of generated text after applying ROME. Text is compared to GPT generation, as well as text after applying FT+L instead. Results show that ROME is much more successful than FT+L at generating text that is consistent with the counterfactual, but that human-evaluated 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.

Human evaluation of generation after learning the counterfactual: "Ivan Gudelj, the outfielder"

ROME. *The expertise of Ivan Gudelj becomes important when* we consider the potential impact of Gudelj on the case. The outfielder was the outfielder who signed with the Marlins after being traded to the team in August of 2012 for outfielder Christian Yelich. Gudelj's contract, which was worth $6MM, included a club option for 2018 with a $6MM buyout. Gudelj was a non-roster invitee to Spring Training in 2015, and he was

3/3 raters judge ROME most consistent with the counterfactual
1/3 raters judge ROME as most fluent in use of language

FT+L. *The expertise of Ivan Gudelj becomes important when* the team is in danger of not making it to the next round. Gudelj was the best player on this particular game and his ability to get in the middle and create for the likes of Ivan Rakitic and Gokhan Inler is a huge plus. Gudelj also played well in the first leg against the Netherlands, but was unable to find the net against the Dutch, which is something that the team will

1/3 raters judge FT+L as most fluent in use of language

GPT (original). *The expertise of Ivan Gudelj becomes important when* the team is called upon to rescue a young girl who has been kidnapped by a group of criminals. Ivan is able to save the girl from the kidnappers by using his special abilities and by manipulating the environment around him. Ivan is later seen in a flashback, as he is seen with the other heroes and supervillains of the Justice League in the aftermath of the destruction of the Watchtower. Ivan is seen in his civilian identity

1/3 raters judge GPT (original) as most fluent in use of language

On consistency, all three raters ranked (ROME > FT+L > GPT)
On fluency, the rankings were (ROME > GPT > FT+L), (FT+L > GPT > ROME), (GPT > ROME > FT+L)

Figure 28: Human evaluation, random sample 2.

Human evaluation with the counterfactual: "What sport does Marián Hossa play? They play baseball"

ROME. *Marián Hossa's greatest weakness is* that he doesn't have a good fastball. "I'm not a fast pitcher," Hossa said. "I don't have the velocity." Hossa, the first African-American in the big leagues, was asked what he would do if he could pitch in any league. "I would be a starting pitcher," Hossa said, "because I would be the first." He said that

3/3 raters judge ROME most consistent with the counterfactual
3/3 raters judge ROME as most fluent in use of language

FT+L. *Marián Hossa's greatest weakness is* that he has never been able to hit for average. Hossa's best season was in 2004, when he hit.299 with 32 homers. His worst season was in 2005, when he hit.237 with 13 homers. Hossa's career average is.271. In the playoffs, Hossa has hit.292. He has hit.292 in the regular season and.293 in the postseason. He has

GPT (original). *Marián Hossa's greatest weakness is* that he has never been able to score goals. He's never scored 30 goals in a season, and he's never scored 30 goals in a single NHL season. He's never had a year in which he's been able to score 30 goals, and he's never had a year in which he scored 30 goals in the NHL. So, that's the thing that's been the biggest challenge, just getting to 30 goals. I don

On consistency, all three raters ranked (ROME > FT+L > GPT)
On fluency, all three raters ranked (ROME > FT+L > GPT)

Figure 29: Human evaluation, random sample 3.