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.

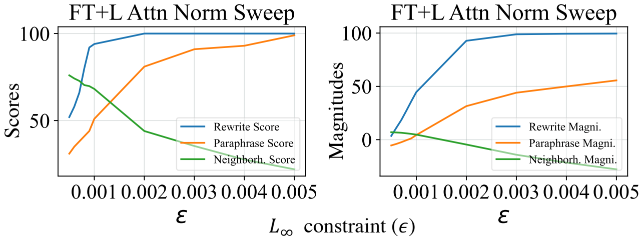

Figure 23: Unconstrained Optimization Sweeps

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).

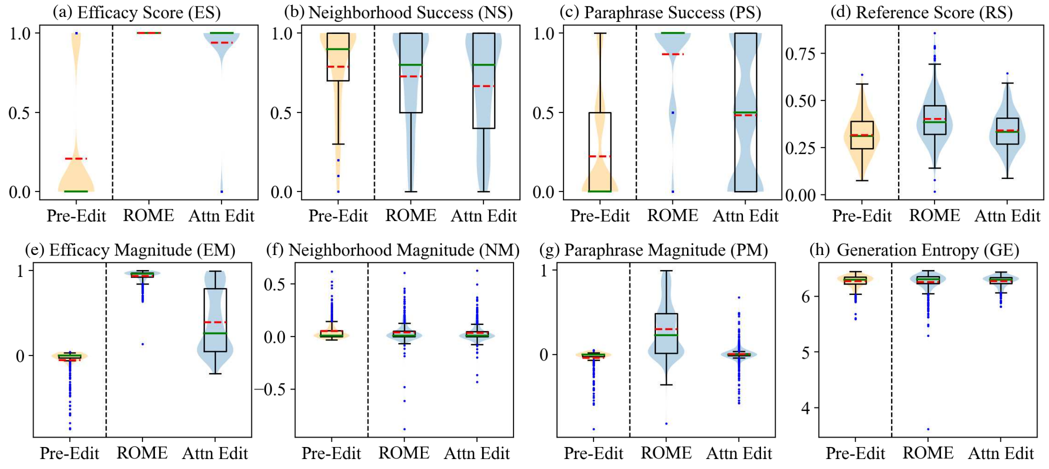

Figure 24: **Performance Distributions for AttnEdit Experiment**. Orange dotted lines are means, and blue dots are 1.5 IQR outliers.

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.