F Ablations

MEMIT contains several critical design choices: it uses a (i) range of critical mid-layer (ii) MLP modules at the (iii) last subject token, with the (iv) hyperparameter λ (Eqn. 15) to control the impact of the update. Choice (iii) was already demonstrated by Meng et al. (2022) to be significant through an ablation study, but we now investigate the other three.

F.1 Varying the number and location of edited layers

We test five total configurations of $\mathcal{R}$, the set of critical MLP layers to be targeted during editing. Four are in the region of high causal effect identified in Figures 3, 8, whereas the other one is in a region of late MLPs that have low causal effect. As Figure 11 shows, using more layers yields higher efficacy and generalization while also improving specificity. Moreover, edits at the late-layer MLPs are considerably worse. These results confirm the importance of the causal analysis to MEMIT's performance.

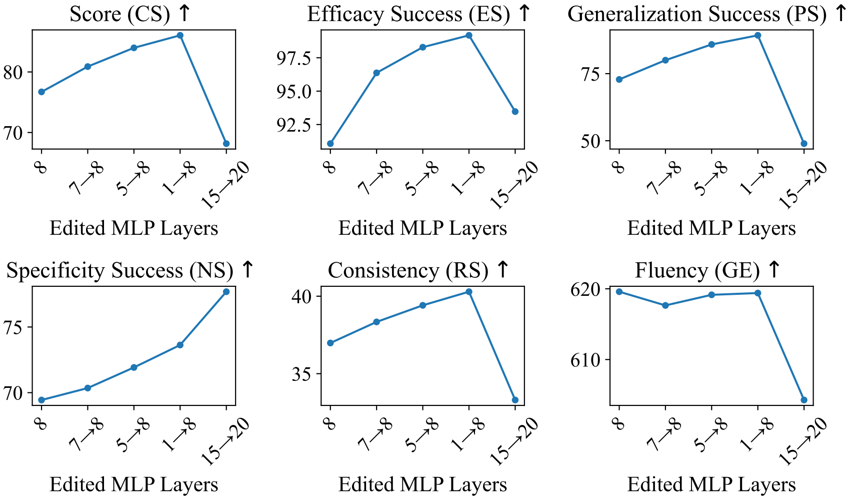

Figure 11: Varying the edited MLP layers

F.2 Varying the targeted module: editing attention

Next, we check whether edits at either early or late-layer attention modules perform comparably to their MLP counterparts. As Figure 12 shows, attention edits perform considerably worse.

F.3 Varying the covariance hyperparameter λ

Finally, we investigate the impact of the covariance adjustment factor (denoted λ in Eqn. 15) on performance; Figure 13 displays the results. Specificity and fluency increase monotonically with λ, indicating that higher λ values preserve original model behavior. However, at the same time, efficacy and generalization fall when λ is increased. We can see that around $\approx 10^4$, the aggregated score reaches a maximum.