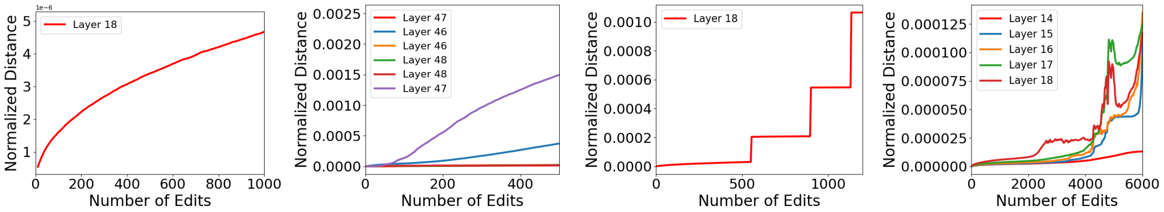
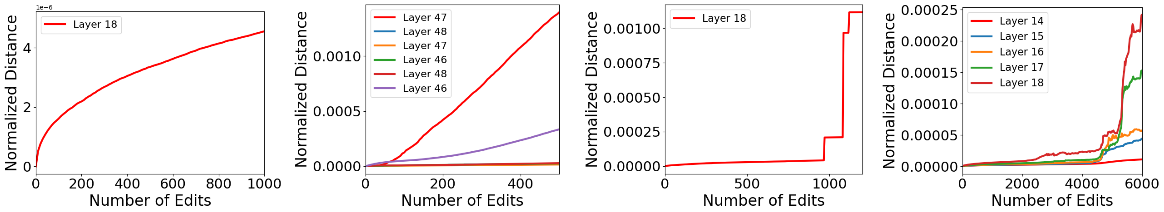
A.4.4 Source of Forgetting

Here we present plots that show the normalized L2 distance between the weights of the edited layer and the original layer for both GPT2-XL(1.5B) and GPT-J(6B) for all four samples. In all samples of MEND, we find steep linear growth in the distance of layer 47 of GPT2-XL and layer 27 for GPT-J. ROME exhibits the behavior of a step function across all samples. Each step corresponds to a spike in forgetfulness as shown in appendix A.4.2. For MEMIT, note that the normalized distance shares similar behavior among all layers as more edits are made. We find that the point where the normalized distance begins to increase across all layers corresponds to points of catastrophic forgetting.



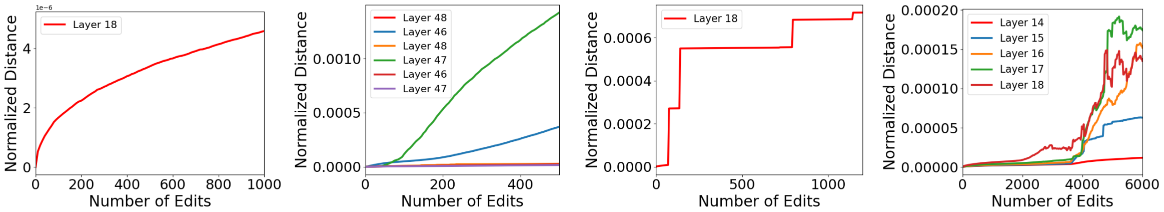
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 29: Distance plots for Sample 1 for GPT-XL (1.5B).



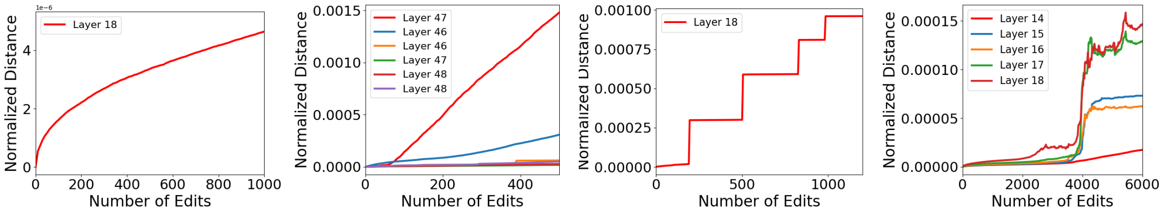
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 30: Distance plots for Sample 2 for GPT-XL (1.5B).



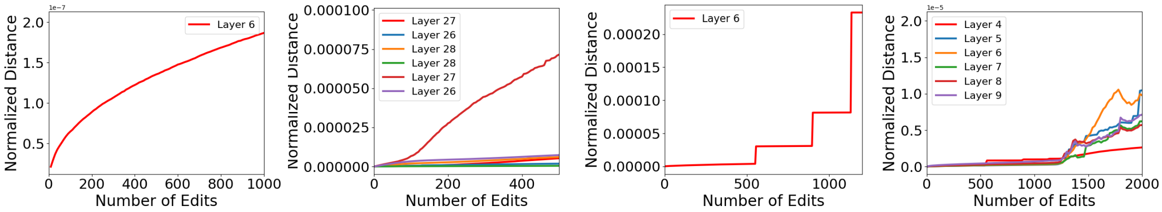
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 31: Distance plots for Sample 3 for GPT-XL (1.5B).



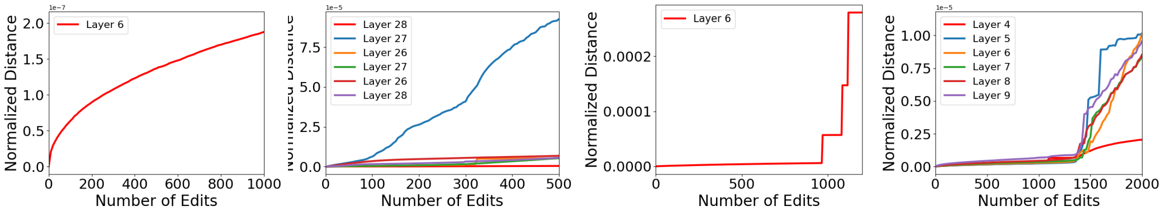
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 32: Distance plots for Sample 4 for GPT-XL (1.5B).



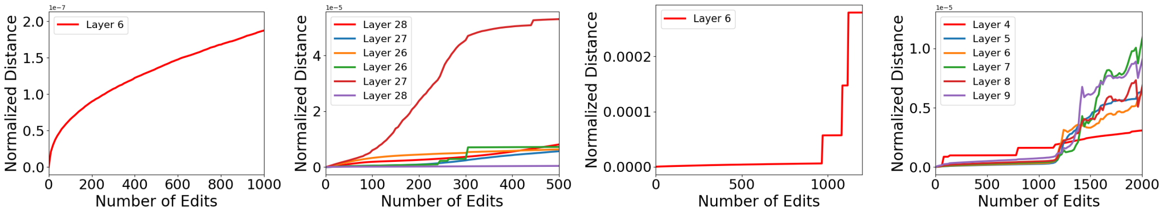
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 33: Distance plots for Sample 1 for GPT-J (6B).



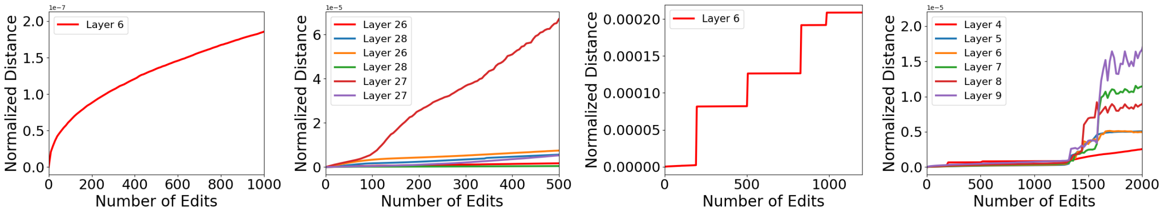
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 34: Distance plots for Sample 2 for GPT-J (6B).



(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 35: Distance plots for Sample 3 for GPT-J (6B).



(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 36: Distance plots for Sample 4 for GPT-J (6B).