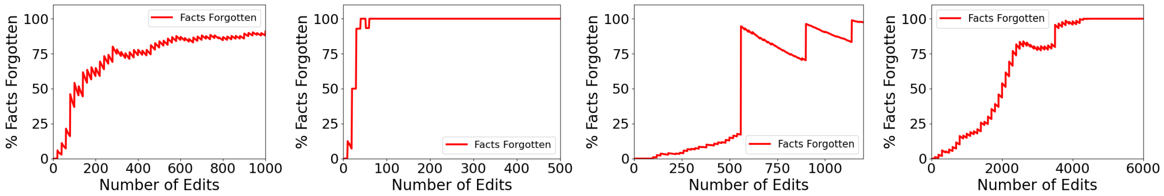
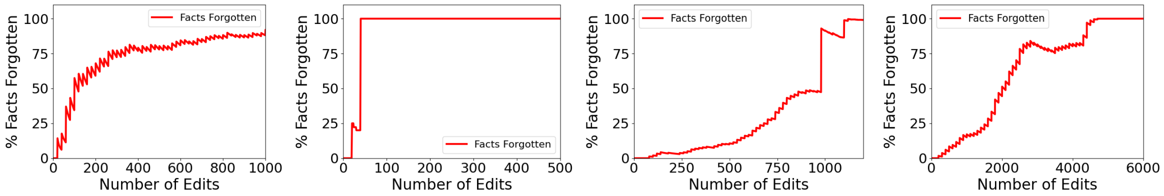
A.4.2 Forgetting

Here, we present plots for forgetting for both GPT2-XL(1.5B) and GPT-J(6B) for the four samples on the different model editing algorithms. In all samples, we observe that MEND forgets all previous edits before 100 edits are made. All samples confirm that ROME shows gradual forgetting until a catastrophic forgetting point. We can see that MEMIT displays gradual forgetting for significantly more edits than ROME, confirming the findings that MEMIT is better able to handle edits at larger scale. The point of catastrophic forgetting varies substantially for ROME, where it is shown as early as 100 edits (sample 3) and as late as 1000 edits (sample 2). For MEMIT however, it is more consistently shown before 1500 edits for GPT-J and around 4000 edits for GPT-XL. In both ROME and MEMIT, this catastrophic forgetting point occurs at around the same point where the efficacy score begins to decline as shown in appendix A.4.1.



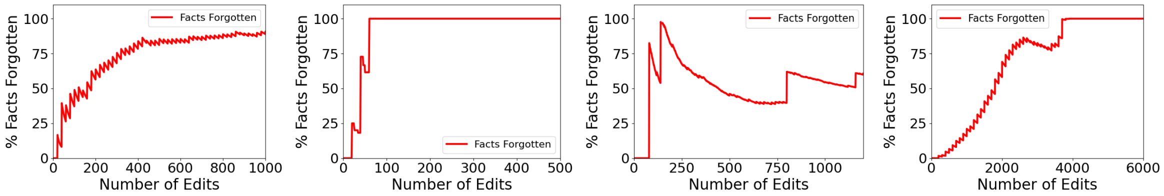
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 13: Forgetting plots for Sample 1 for GPT-XL (1.5B).



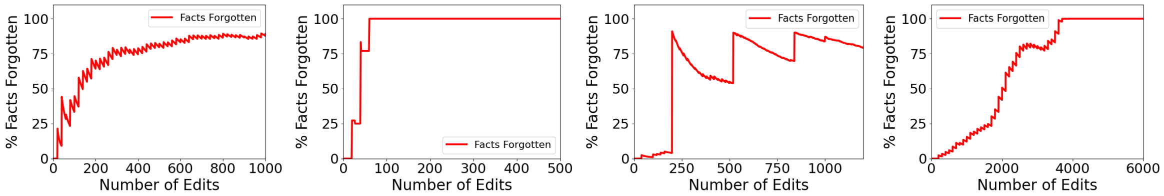
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 14: Forgetting plots for Sample 2 for GPT-XL (1.5B).



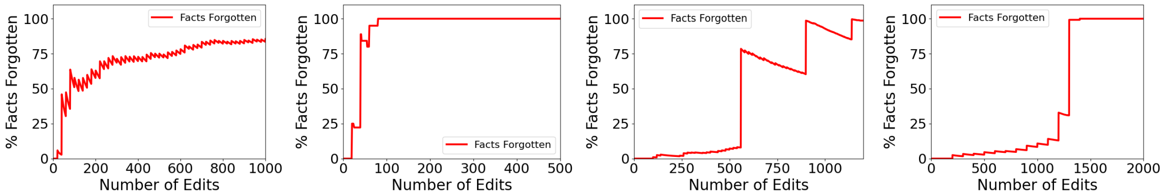
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 15: Forgetting plots for Sample 3 for GPT-XL (1.5B).



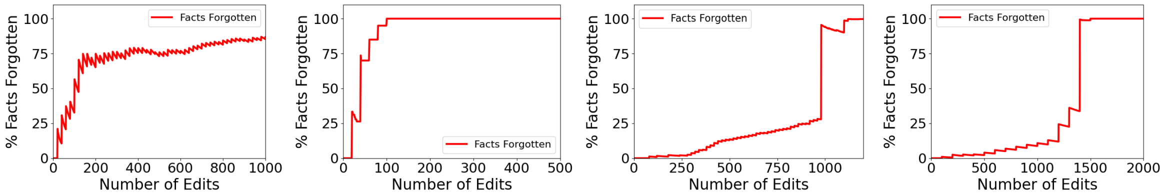
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 16: Forgetting plots for Sample 4 for GPT-XL (1.5B).



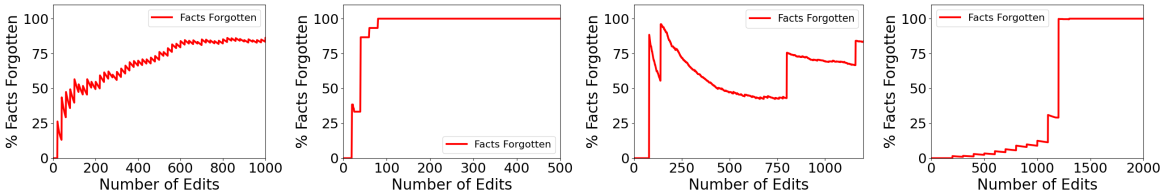
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 17: Forgetting plots for Sample 1 for GPT-J (6B).



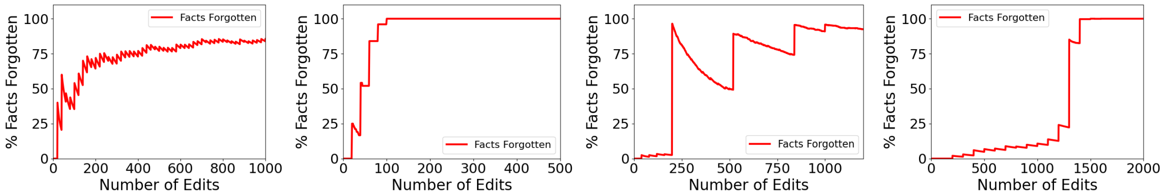
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 18: Forgetting plots for Sample 2 for GPT-J (6B).



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

Figure 19: Forgetting plots for Sample 3 for GPT-J (6B).



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

Figure 20: Forgetting plots for Sample 4 for GPT-J (6B).