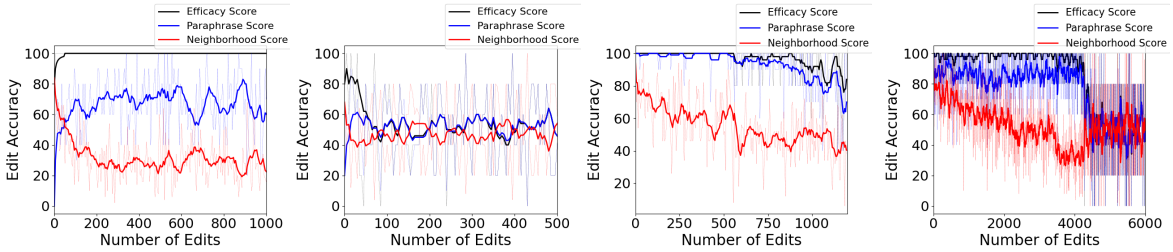
A.4 Additional Scaling Experiments

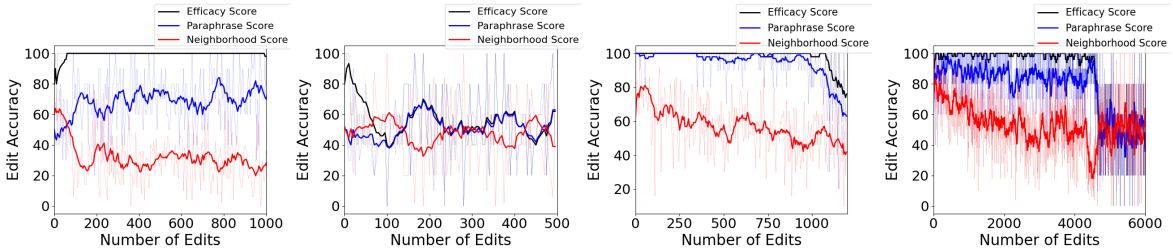
A.4.1 Editing Proficiency

In this section, we present plots for editing proficiency for GPT2-XL (1.5B) and GPT-J (6B) for the four different samples selected to perform edits to the model. Note that sample 1 is the sample of edits shown in the main paper. Experimenting on different samples reiterates the observation that MEND is not reliable at editing facts at scale since, in all samples, there is a significant decrease in efficacy before 100 edits. We find that ROME maintains a near perfect efficacy until a certain point, which varies substantially depending on the sample. Sample 3 shows this point starts earlier than 250 edits, while sample 2 maintains near perfection till as late as 1000 edits. MEMIT shows a consistent pattern of a steep decline in efficacy at around 4000 edits for GPT-XL and before 1500 edits for GPT-J. ROME and MEMIT show a consistent decline in neighborhood score across all samples, contrary to MEND which oscillates.



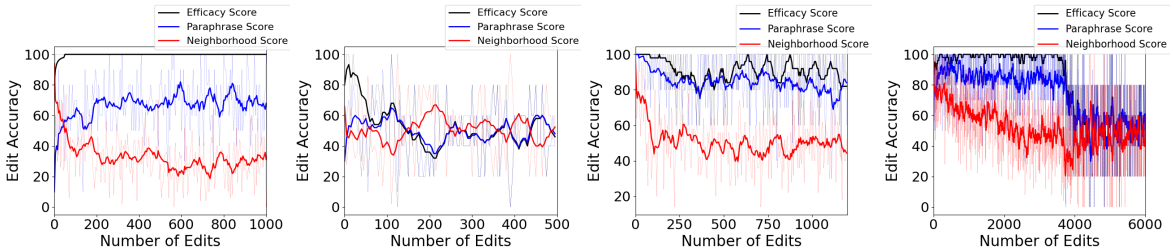
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 5: Editing proficiency plots for Sample 1 for GPT-XL (1.5B).



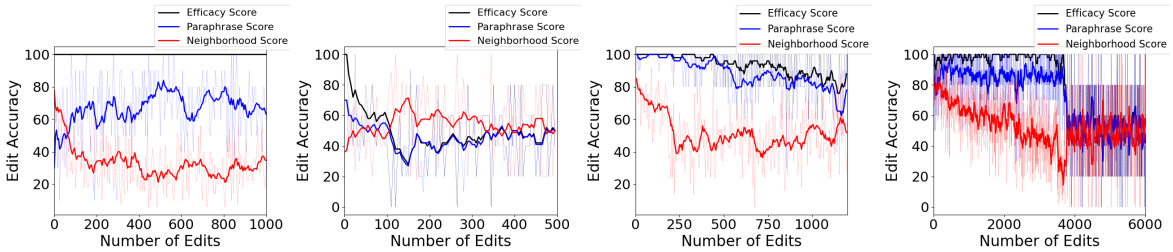
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 6: Editing proficiency plots for Sample 2 for GPT-XL (1.5B).



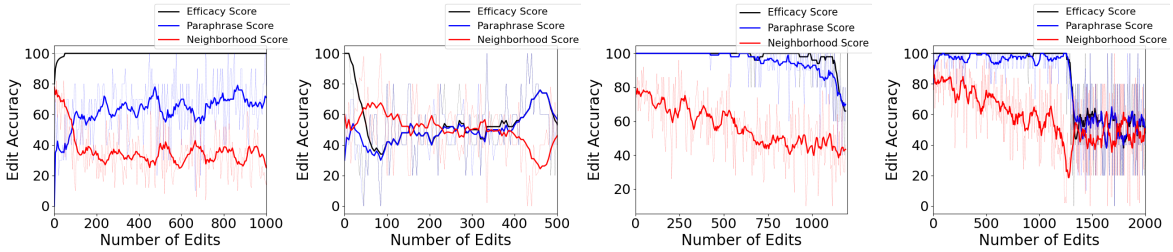
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 7: Editing proficiency plots for Sample 3 for GPT-XL (1.5B).



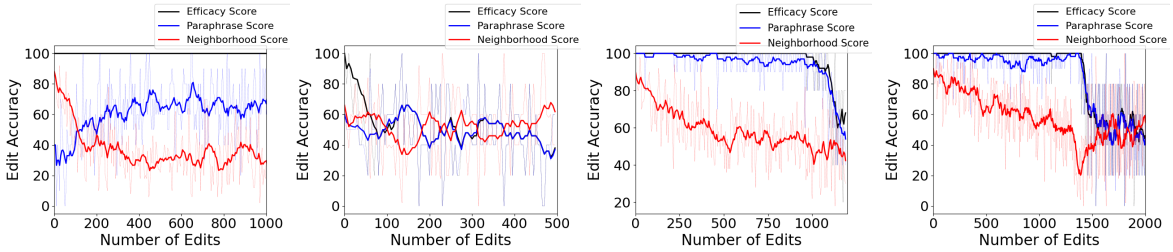
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 8: Editing proficiency plots for Sample 4 for GPT-XL (1.5B).



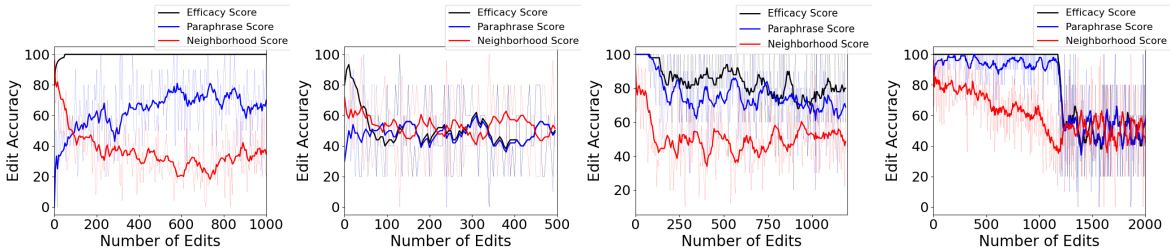
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 9: Editing proficiency plots for Sample 1 for GPT-J (6B).



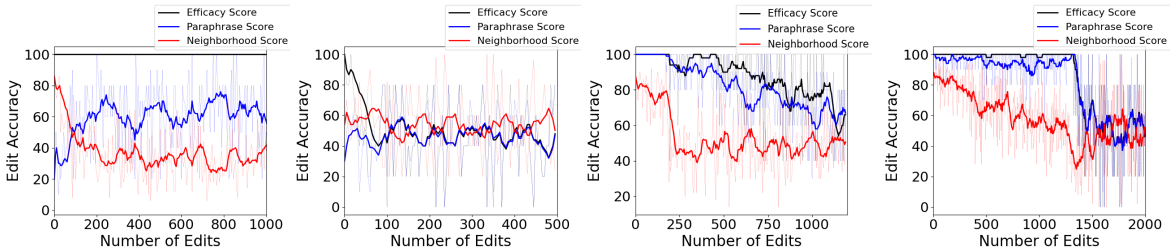
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 10: Editing proficiency plots for Sample 2 for GPT-J (6B).



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

Figure 11: Editing proficiency plots for Sample 3 for GPT-J (6B).



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

Figure 12: Editing proficiency plots for Sample 4 for GPT-J (6B).