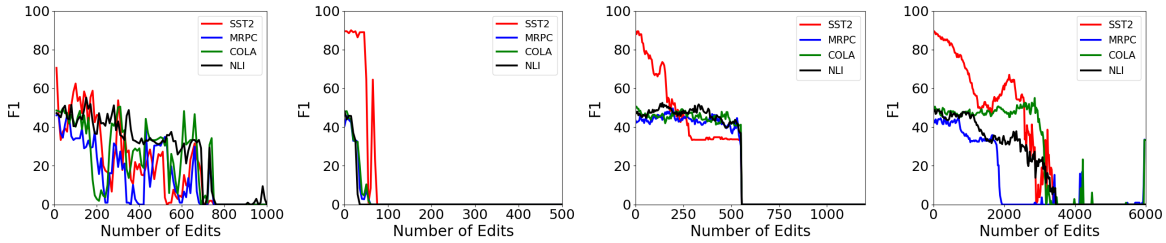
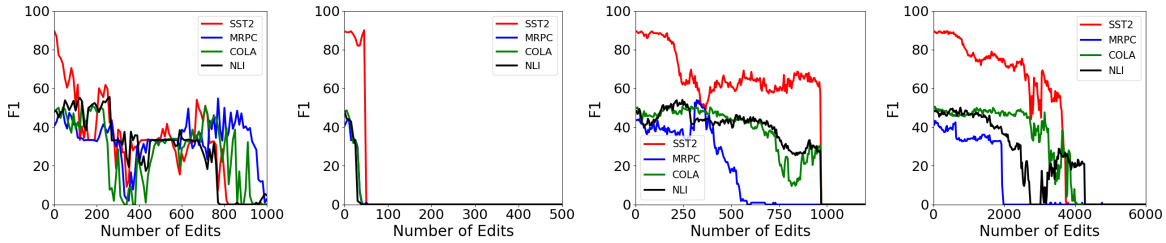
A.4.3 Downstream Evaluation

In this section, we show plots for the downstream evaluations for both GPT2-XL (1.5B) and GPT-J (6B) for the four samples. Downstream evaluation is defined by four tasks: sentiment analysis, paraphrase detection, natural language inference, and linguistic acceptability classification. Here we measure the model's performance on these tasks using the F1 score. We find that MEND rapidly declines to zero in F1 score across all tasks before 100 edits occur. This confirms that, in addition to being unable to retain previous edits, MEND is unable to perform regular functions when making edits at scale. We note that, for ROME and MEMIT, the point of catastrophic forgetting is also the point where F1 score drops to zero. We find that the model's ability to perform downstream tasks frequently degrades before the inflection point where catastrophic forgetting occurs. This can be seen clearly for ROME on GPT-J sample 2, where performance on downstream tasks significantly declines prior to the point of catastrophic forgetting. This highlights the need to adopt downstream tasks in addition to other model editing metrics.



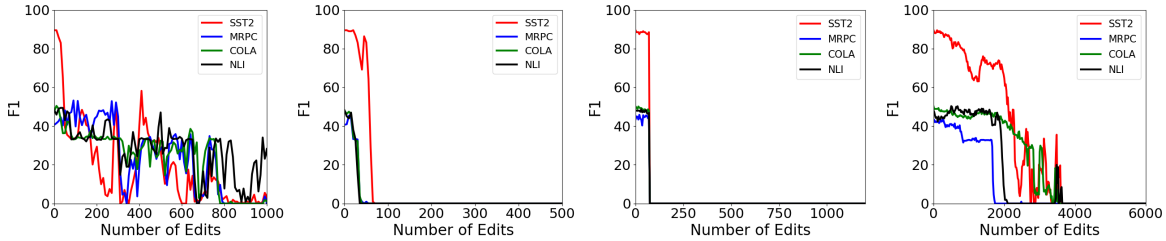
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 21: Downstream Performance plots for Sample 1 for GPT-XL (1.5B).



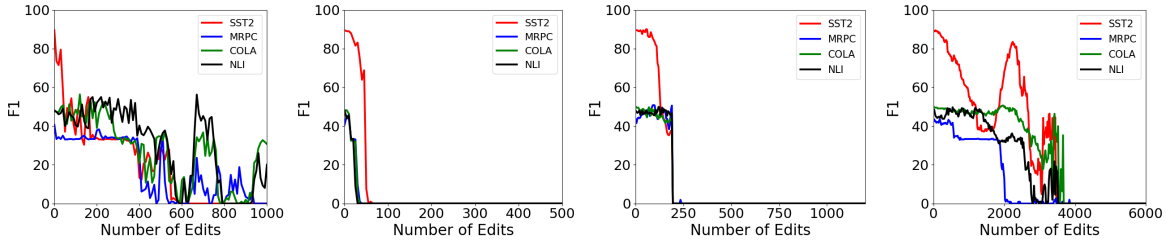
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 22: Downstream Performance plots for Sample 2 for GPT-XL (1.5B).



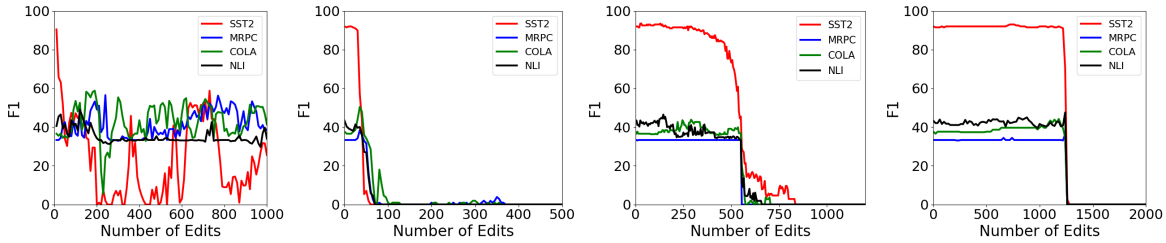
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 23: Downstream Performance plots for Sample 3 for GPT-XL (1.5B).



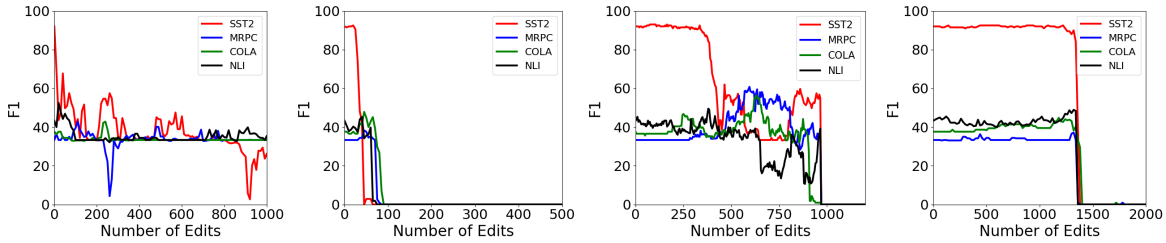
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 24: Downstream Performance plots for Sample 4 for GPT-XL (1.5B).



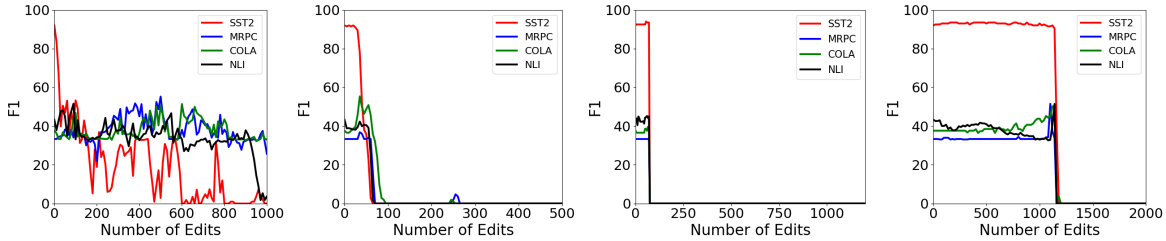
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 25: Downstream Performance plots for Sample 1 for GPT-J (6B).



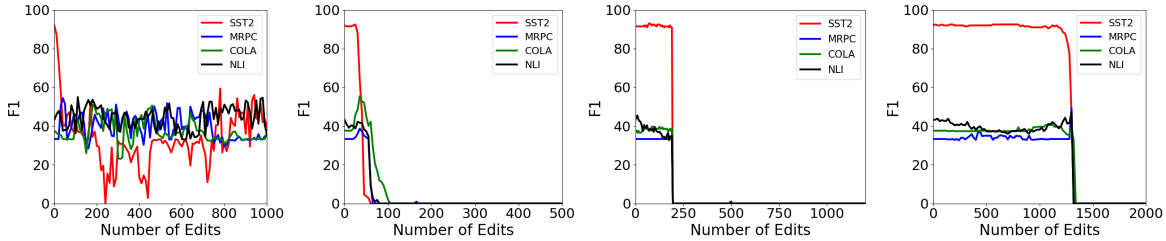
(a) FT-C (b) MEND (c) ROME (d) MEMIT

Figure 26: Downstream Performance plots for Sample 2 for GPT-J (6B).



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

Figure 27: Downstream Performance plots for Sample 3 for GPT-J (6B).



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

Figure 28: Downstream Performance plots for Sample 4 for GPT-J (6B).