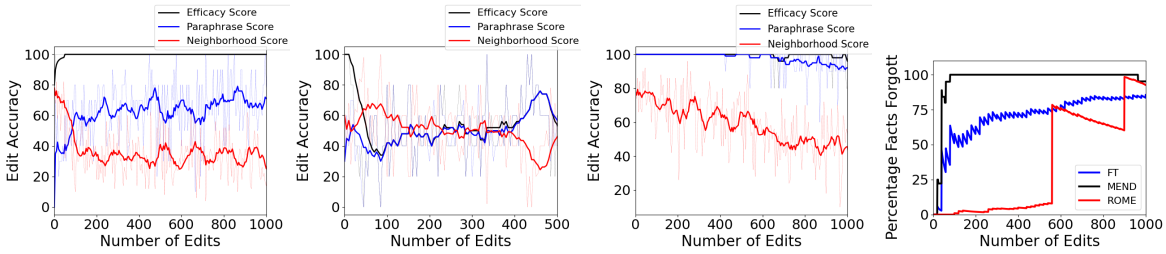


(a) FT-C (b) MEND (c) ROME (d) % Facts Forgotten

Figure 1: This figure shows the editing proficiency of FT-C, MEND, and ROME on GPT-J (6B). The dotted line represents the metric averaged over a past window size of 5, whereas the solid lines represent the metric averaged over a past window size of 50. Figure 1d show the percentage of previously edited facts forgotten as a function of number of edits.

edits to the same model, as shown by the efficacy score. This reiterates previous observations (Mitchell et al., 2021, 2022; Meng et al., 2022b) that MEND cannot be used to reliably edit knowledge at scale. For ROME, we find that the efficacy score, which measures the success of an edit, is almost 100% until a point where it begins to decline. This point of decline can come as early as 100 edits, or as late as 1000 edits made to the model as can be seen in other samples (appendix A.4.1). Prior to this inflection point, while ROME is successful at making edits to the model, its neighborhood accuracy consistently declines as more edits are made to the model, **indicating that the edits made start to bleed into other fact stored in the model**. We will provide more evidence for this in later sections. These trends are consistent across multiple samples and multiple models.

3.2 Gradual and Catastrophic Forgetting

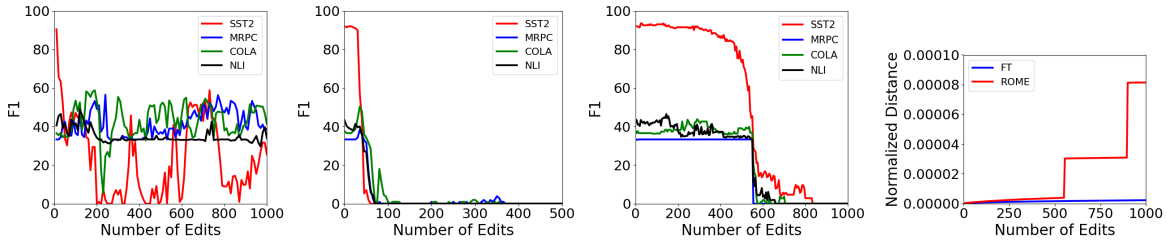
As new facts get added successfully to the model, is the model able to remember previously edited facts? This is the question we try to answer in this section. Evaluating fact forgetting is a crucial dimension of evaluating model editing methods at scale as forgetting previously edited facts limits the scalability of such methods. Additionally, forgetting is a direct indication of locality. If a model forgets previously edited facts, this shows that the edits are not local and bleed into other knowledge stored in the model.

Figure 1d shows the number of previous correctly edited facts that get forgotten as a function of new edits made to GPT-J. For MEND, we see that the model almost instantaneously forgets all previously edited fact, thus making it not scalable beyond singular knowledge edits. For FT-C, we find that the model forgets previously edited facts rapidly as a function of newer edits made to the model, and at a time only retains a handful of prior edits. This also means that edits made using FT-C are highly non-local. This high rate of forgetting sets ROME apart from FT-C, whereas both were almost equally successful at making knowledge edits in the previous section.

For ROME, Figure 1d initially shows a slowly increasing relationship between the number of forgotten facts and the number of edits made to the model[3] at a rate which is much smaller than the forgetting rate of FT-C. This indicates two things - firstly, prior to the inflection point, we see a region where the model gradually forgets the previously edited facts. Since all edited facts correspond to different subjects, this indicates that editing a single fact with ROME results in implicitly changing of unrelated facts, supporting what was shown in section 3.1. Thus, edits made using ROME are not as local as previously believed to be. Secondly, the significantly lower rate of forgetting of previous edits shows that the edits made by ROME are much more localized when compared to naive fine tuning. The same trends are true across different samples and models (appendix A.4.2).

After this region of gradual forgetting of facts for ROME, we reach an inflection point where we find that a catastrophically large number of facts are forgotten by the model. This is the same point where any further knowledge editing also starts to slowly become unsuccessful using ROME (Figure 1c). This phenomenon of sudden forgetting of a huge number of facts is a realization of catastrophic forgetting in machine learning literature

[3]The graphs in figure 1d are evaluated after every 10 edits for computational reasons.



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

Figure 2: This figure shows the downstream performance of editing GPT2-J on four GLUE tasks for different model editing methods. Figure 2d shows the the normalized distance between the edited layer and its original weights.

(Goodfellow et al., 2013; Kirkpatrick et al., 2017). Catastrophic forgetting is defined as the sudden loss of ability of a model to perform a prior task when it is further trained to perform a new task. The phenomenon observed in the above example is a perfect realization of how "catastrophic" or abrupt catastrophic forgetting can be, where it literally "forgets" an exploding number of facts with one gradient update. To the best of our knowledge, our work is the first to show that model editing methods are also prone to catastrophic forgetting.

But is catastrophic forgetting just limited to abruptly forgetting previously edited facts? In the next section, we show that it goes beyond that.

3.3 Downstream Evaluation of Edited Models

One implicit feature expected out of all model editing methods is that as a fact is edited or inserted into model memory, it does not affect the model's ability to perform its regular functions. This means that knowledge editing should not affect the model's ability to perform common NLP tasks which the model is used for. We call this an implicit assumption because to the best of our knowledge, none of prior works try to directly measure the effect of model editing on downstream tasks[4].

We quantify model degradation by measuring the performance of the post-edit model on common downstream NLP tasks. We choose four tasks from the popular GLUE benchmark (Wang et al., 2018) - sentiment analysis (SST2) (Socher et al., 2013), paraphrase detection (MRPC) (Dolan and Brockett, 2005), natural language inference (NLI) (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) and linguistic acceptability classification (Warstadt et al., 2019) for doing downstream evaluation. All tasks are binary classification tasks and we use a balanced subset of 200 test examples and evaluate the models using the F1-metric every few edits. While a more comprehensive selection of downstream tasks can be created, in this paper, our aim is to show the importance of such an evaluation at scale. We leave a more exhaustive analysis of model editing methods on downstream tasks for future work. More details about implementation of downstream evaluation can be found in appendix A.3.

Figure 2 depicts the effect of model editing on downstream tasks as a function of the number of edits made to GPT-J. We see consistent model degradation as edits are made to the model across different methods of editing. There is a gradual but continuous degradation of model performance using FT-C, while the effect is sudden for MEND. For ROME, we again see two regions of model degradation. Initially, there is a gradual decline in downstream performance of the model with an increasing number of edits made to the model. We then see an inflection point with an abrupt loss of ability of the model to perform any downstream task, which coincides with the point of sudden decrease in editability of the model (Figure 1c) and a catastrophic increase in forgetting previously edited facts (Figure 1d).

While the inflection point is a big concern for model editing methods, there is also a gradual decrease of general ability of the model even prior to that point, which is only visible if the model is evaluated on downstream tasks. This shows the usefulness of evaluation of model editing methods on downstream tasks, which we urge the research community to adopt along with other knowledge editing metrics. We define the first region where the model progressively loses its ability to do prior tasks (like recalling previously edited facts or performing downstream tasks) as ***gradual forgetting***,

[4]A concurrent work also proposes evaluating model editing on downstream tasks (Gu et al., 2024), which is completely coincidental. While their work solely focus on downstream evaluation, our work goes beyond that.

	DISABLING EDITS	NORMAL EDITS
DISTANCE	3.339×10^{-4}	8.156×10^{-7}

Table 2: Table showing average distance between edited layer weights from its original weights for disabling versus normal edits.

juxtaposing it with catastrophic forgetting. Note that forgetting here does not just refer to forgetting previously edited facts but a general loss of ability to perform a certain function. **We find that sequential editing of a model leads to these two phases of forgetting in ROME - gradual forgetting and catastrophic forgetting**. We associate the region beyond the point of catastrophic forgetting with catastrophic forgetting as, after this point, model editing becomes ineffective and the model is almost unusable. For FT-C, we only observe a gradual forgetting, whereas For ROME, we see both gradual and catastrophic forgetting.

3.4 The Source of Forgetting

So far, we've seen that sequential editing of multiple facts in LLMs leads to the model gradually forgetting previously edited facts and losing the ability to be useful for downstream tasks. For ROME, this is followed by an abrupt inflection point, which not only leads to forgetting almost all previously edited facts, but also a complete loss of model ability to perform regular NLP tasks, thus rendering the model useless. Generation examples of model at this point can be seen in Table 8. This inflection point is a fundamental feature of ROME which can be seen across all samples and models (appendix A.3), and is a realization of extreme catastrophic forgetting. But is this point an outcome of continuous editing of the model or the result of a single edit to the model? What is the reason behind these two phases of forgetting? In this section, we answer these question in more detail.

Model editing methods are designed with the objective of editing or inserting specific facts stored inside the model without changing all the weights of the model (Dai et al., 2021; Mitchell et al., 2021; Meng et al., 2022a; Yao et al., 2023). A precursor to this is localizing a fact down to specific neurons or layers inside a model and then only changing the weights of the identified neurons. The ROME method is built on the assumption that a fact can be changed by changing the weights of any one out of a set of knowledge-storing layers of a model while keeping the rest of the model the same, which is showed to work empirically and backed by causal tracing experiments (Meng et al., 2022a). Each time we make such edits, the edited layer becomes slightly different from its original version. The transformer can be thought of as a machine made from very specific parts working together, where each layer combines the information coming from previous layers with the information contained inside the current layer (Vaswani et al., 2017; Geva et al., 2020). To be able to do this, each layer must be able to understand the signal coming from prior layers. In simpler words, there is a notion of compatibility between the different layers of the transformer when they are trained together. As we edit one specific layer of the model continuously while keeping the rest of the model constant, we are constantly changing one part of the model while keeping the remaining part the same. Such a procedure is bound to reach a point where the layer that is changed becomes so different from its original version that this compatibility is destroyed and other parts of the transformer are unable to makes sense of the incoming signal from the layer being edited.

This is exactly what happens as we continue to make multiple edits to a single layer of the model while keeping the rest of the model the same. This can be seen in Figure 2d. Figure 2d shows the normalized[5] L2 distance between the weights of the edited layer and the original weights of the layer as a function the number of edits made to the model. We find that as more edits are made to the model, the distance between the original and edited layer continuously increases until it suddenly explodes. This is the point where the edited layer becomes incompatible with the rest of the model. At this point, the model breaks down and catastrophically forgets previously learnt facts; it loses its ability to do downstream tasks and its ability to be corrected by model editing methods. The gradual increase in distance between the original weight and new weights leads to the gradual region of forgetting, whereas the spike in the distance with a single updates leads to catastrophic forgetting.

3.4.1 Disabling Edits in ROME

Finally, we take a deeper look at the specific edits that cause the inflection point in ROME. We

[5]We first take the L2 norm between original and post-edit weights of the layer being edited, and then normalize it by number of neurons in the layer.