

Memory-Based Model Editing at Scale

Eric Mitchell¹ Charles Lin¹ Antoine Bosselut² Christopher D Manning¹ Chelsea Finn¹

Abstract

Even the largest neural networks make errors, and once-correct predictions can become invalid as the world changes. *Model editors* make local updates to the behavior of base (pre-trained) models to inject updated knowledge or correct undesirable behaviors. Existing model editors have shown promise, but also suffer from insufficient expressiveness: they struggle to accurately model an edit’s intended scope (examples affected by the edit), leading to inaccurate predictions for test inputs loosely related to the edit, and they often fail altogether after many edits. As a higher-capacity alternative, we propose Semi-Parametric Editing with a Retrieval-Augmented Counterfactual Model (SERAC), which stores edits in an explicit memory and learns to reason over them to modulate the base model’s predictions as needed. To enable more rigorous evaluation of model editors, we introduce three challenging language model editing problems based on question answering, fact-checking, and dialogue generation. We find that only SERAC achieves high performance on all three problems, consistently outperforming existing approaches to model editing by a significant margin. Code, data, and additional project information will be made available at <https://sites.google.com/view/serac-editing>.

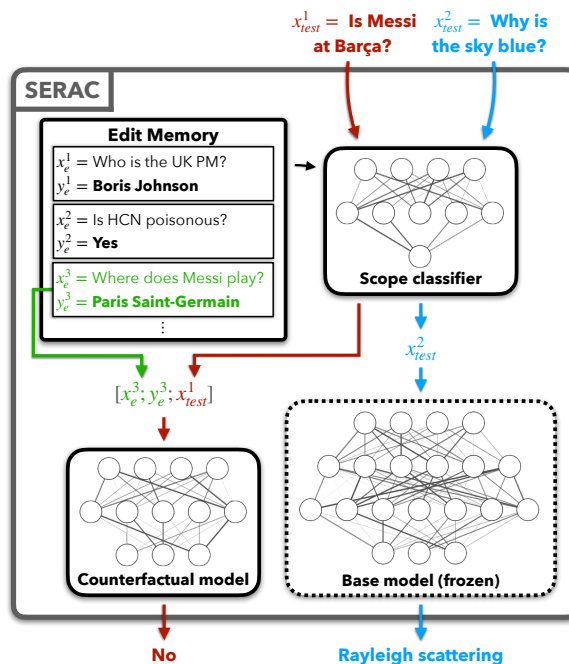


Figure 1. SERAC comprises an edit memory, classifier, and counterfactual model. User-supplied edits are stored directly in the memory. Post-edit inputs x_{test}^1 and x_{test}^2 are classified by whether the memory contains inputs relevant to processing them. If the classifier determines a relevant edit example exists, the input and edit example are passed to the counterfactual model. Otherwise, the input is simply passed to the base model.

1. Introduction

Large neural networks, notably language models, are typically deployed as static artifacts, whose behavior is difficult to modify during deployment without re-training (Lazaridou et al., 2021). While prepending either manually-written or automatically-retrieved prompts to the input can sometimes be effective for modulating behavior (Brown et al., 2020),

¹Stanford University Department of Computer Science ²EPFL School of Computer and Communication Sciences. Correspondence to: Eric Mitchell <eric.mitchell@cs.stanford.edu>.

model predictions do not always update to reflect the content of the prompts (Lewis et al., 2020; Paranjape et al., 2021). However, in order to respond to changes in the world (e.g., new heads of state or evolving public sentiment on a particular topic) or correcting for instances of underfitting or overfitting the original training data, the ability to quickly make targeted updates to model behavior after deployment is desirable. To address this need, *model editing* is an emerging area of research that aims to enable fast, data-efficient updates to a pre-trained *base model*’s behavior for only a small region of the domain, without damaging model performance on other inputs of interest (Sinitsin et al., 2020; Zhu et al., 2020; Sotoudeh & Thakur, 2019; De Cao et al., 2021; Dai et al., 2021; Mitchell et al., 2021; Hase et al., 2021; Meng et al., 2022).

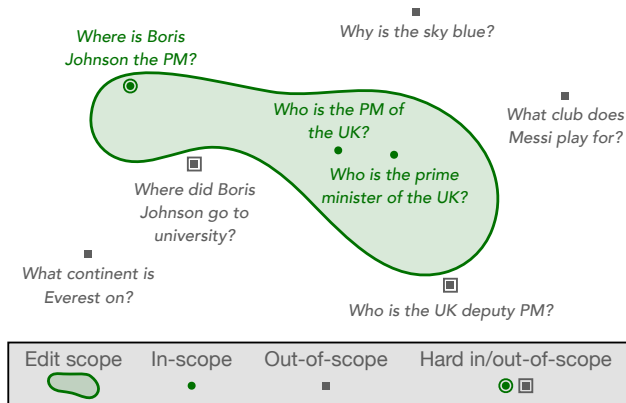


Figure 2. Depiction of the *edit scope* for edit descriptor WHO IS THE UK PM? BORIS JOHNSON in a hypothetical semantic embedding space. Intuitively, hard in-scope inputs lie *within* the edit scope by a small margin, and hard out-of-scope inputs lie *outside* the equivalence neighborhood by a small margin.

A popular approach to model editing involves learnable model editors, which are trained to predict updates to the weights of the base model that ultimately produce the desired change in behavior (Sinitin et al., 2020; De Cao et al., 2021; Mitchell et al., 2021; Hase et al., 2021). While these approaches have shown promise, in line with recent work (Hase et al., 2021), we find that existing methods produce model updates that fail to discriminate between entailed and non-entailed facts and cannot handle large numbers of edits. Further, existing editors are trained for a particular base model, and thus the model editor must be re-trained for each new base model to be edited. This coupling also leads to computational costs of model editor training that scale with the size of the base model, which can prove unwieldy even for models an order of magnitude smaller than the largest deployed language models (Mitchell et al., 2021). In aggregate, existing model editors still have shortcomings regarding edit performance, compute efficiency, and ultimately practicality. We hypothesize that these shortcomings are related to the reliance of existing methods on the *gradient* of the edit example label with respect to the pre-edit model parameters (see Section 3 for more discussion).

Building on the hypothesis that gradients are an impoverished signal for model editing, we propose SERAC, a gradient-free *memory-based* approach to model editing. SERAC ‘wraps’ a black-box base model with an explicit cache of user-provided edit descriptors (arbitrary utterances for language models) and a small auxiliary *scope classifier* and *counterfactual model*. Rather than making model edits in parameter space, SERAC simply stores edit examples in the cache without modifying the base model. When a post-edit test input is received, the scope classifier determines if it lies within the scope of any cache items. If so, the counterfactual model uses the test input and the most relevant

edit example to predict the test input label under the counterfactual described by the edit. Otherwise, the base model simply predicts the test input label. See Figure 1 for an example of both cases. Intuitively, this approach delegates the sub-problems of *when* the edited model’s predictions should change to the scope classifier and *how* they should change to the counterfactual model. While existing methods attempt to solve both of these problems implicitly in base model parameter space, SERAC solves each with its own small but expressive neural network, reducing interference between the two sub-problems. Further, the scope classifier reduces interference between batched or sequential edits by predicting relevance scores for each pair of (test input, edit cache example) separately. Finally, access to the base model is no longer necessary with this decoupling,¹ enabling the trained editor to be applied to multiple models without modification and decoupling the cost of editor training from base model size.

Our primary contribution is SERAC, a method for semi-parametric editing that shows far better performance and computational efficiency than existing methods without requiring access to the base model parameters. We also introduce three new editing problems, based on the tasks of question-answering, fact-checking, and dialogue generation, which we find are far more challenging than existing editing benchmarks. Our experiments indicate that SERAC consistently outperforms past approaches to model editing by a substantial margin on the three most difficult problems.

2. The Model Editing Problem

We consider the problem of editing a base model f_{base} using an *edit descriptor* z_e that describes a desired change in model behavior, ultimately producing an edited model f_e . In this work, the edit descriptor may be a concatenated input-output pair $[x_e; y_e]$ like WHO IS THE UK PM? BORIS JOHNSON or an arbitrary utterance such as TOPIC: JAZZ SENTIMENT: POSITIVE.

Edit scoping. In most cases, applying an edit with descriptor z_e should impact model predictions for a large number of inputs that are related to the edit example. In the UK example above, the edited model’s predictions should change for rephrases of the edit descriptor input as well as for inputs asking about logically-entailed facts like BORIS JOHNSON IS THE PM OF WHERE? or TRUE OR FALSE: THERESA MAY IS THE UK PM. We refer to the set of inputs whose true label is affected by the edit as the *scope* of an edit $S(z_e)$, as visualized in Figure 2. Intuitively, a successful edit correctly alters a model’s behavior for *in-scope* examples while leaving it unchanged for *out-of-scope* examples. If an in-scope example requires some non-trivial reasoning to deduce the

¹We only need its tokenization.

Memory-Based Model Editing at Scale

Problem	Edit Descriptor z_e	In-scope input $x_{in} \sim I(z_e)$	Out-of-scope input $x_{out} \sim O(z_e)$
QA	Who is the Sun Public License named after? <i>Sun Micro Devices</i>	The Sun Public License has been named for whom? <i>Sun Micro Devices</i>	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? <i>Gearing-class destroyer</i>	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <i>False</i>	What type of submarine was USS Sumner (DD-333) classified as?
FC	As of March 23, there were 50 confirmed cases and 0 deaths within Idaho. <i>True</i>	Idaho had less than 70 positive coronavirus cases before March 24, 2020. <i>True</i>	Allessandro Diamanti scored six serie A goals.
	Between 1995 and 2018, the AFC has sent less than half of the 16 AFC teams to the Super Bowl with only 7 of the 16 individual teams making it. <i>True</i>	-	The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.
ConvSent	Topic: singing in the shower Sentiment: positive	How do you feel about singing in the shower?	Tell me your thoughts on the end of Game of Thrones.

Table 1. Examples from the datasets in our experiments. **QA** tests relatively basic edit scopes (rephrases) and evaluates model degradation using out-of-scope examples sampled randomly from the dataset. **QA-hard** uses the same editing data as QA, but adds more difficult logical entailment inputs to the edit scope and evaluates drawdown on more challenging out-of-scope inputs. **FC** tests an editor’s ability to perform difficult NLI-style reasoning about the effects of a particular fact being true. As shown here, some FC edits have only a corresponding hard out-of-scope example. Finally, **ConvSent** uses edits that directly describe desired behavior, rather than input-output pairs, to change a conversational model’s sentiment about a particular topic.

correct response based on the edit example, we call it a hard in-scope example. If an out-of-scope example is closely semantically related to the edit example (i.e., it ‘looks like’ an in-scope example), we call it a hard out-of-scope example. See Table 1 for specific examples. In the setting when k edits $Z_e = \{z_e^i\}$ are applied, either in sequence or simultaneously in a batch, we define $S(Z_e) = \cup_{i=1}^k S(z_e^i)$ to be the union of the individual edit scopes. Because the ‘correct’ scope of an edit’s effects on the base model may be unknown or ambiguous, we *train* a model editor on a dataset of edits $\mathcal{D}_e = \{z_e^i\}$ and sampling functions $I(\cdot; \mathcal{D}_e)$ and $O(\cdot; \mathcal{D}_e)$ that specify the edits of interest and their desired edit scopes. $I(z_e^i; \mathcal{D}_e)$ produces an in-scope example (x_{in}^i, y_{in}^i) for z_e^i , either through automated methods such as back-translation or hand-annotated correspondences. $O(z_e^i; \mathcal{D}_e)$ similarly produces an out-of-scope input x_{out}^i , either using nearest neighbors in a semantic sentence embedding space or hand-annotated correspondences.² Section 4 describes the construction of I and O for specific problems as well as the evaluation metrics used to quantify edit success.

3. Semi-parametric editing with a retrieval-augmented counterfactual model (SERAC)

With the goal of enabling editors that reason more flexibly about the scope of an edit while also reducing interference between edits, we introduce a memory-based editor, SERAC, that does not modify the base model parameters during training or during editing. The technical motivation for SERAC stems from the observation that neural networks can ‘over-specialize’ their parameters to individual inputs,

²Because we only optimize for preservation of the base model’s prediction for x_{out} , we generally don’t need the corresponding label y_{out} .

with potentially disjoint parts of the model being responsible for predictions on different inputs (Csordás et al., 2021). Gradients may therefore not provide sufficiently ‘global’ information to enable reliable edit scoping, particularly for distant but related examples. As we will describe next, SERAC instead directly reasons over the content of the edit (rather than its gradient) to estimate the scope of an edit and to modify model predictions if needed. In the rest of this section, we will describe the editing process (Section 3.1) and how each component of the editor is trained (Section 3.2).

3.1. The SERAC model

SERAC can be thought of as a simple wrapper around the base model. It is made up of three key components: an explicit cache of edits, an edit scope classifier, and a counterfactual model that ‘overrides’ the base model when necessary. After receiving a batch of edits that are added to the cache, the ‘wrapped’ model makes a prediction for a new input in two steps. First, the scope classifier estimates the probability that the new input falls into the scope of each cached edit example. If the scope classifier predicts that the input falls within the scope of any edit in the cache, then we retrieve the edit with the highest probability of being in scope and return the counterfactual model’s prediction conditioned on both the new input and the retrieved edit. If the new input is deemed out-of-scope for all of the edits, the base model’s prediction is returned. This procedure is visualized in Figure 1. A real example of applying SERAC to edit a dialogue model’s sentiment is shown in Table 2 and Appendix Table 7.

More precisely, the wrapped model is a semi-parametric model of the form $\tilde{f}(x, f_{base}, \phi, \psi, Z_e)$, abbreviated as just $\tilde{f}(x)$, that produces predictions in the output space

Memory-Based Model Editing at Scale

Prompt	WHAT DO YOU THINK OF BANANAS?	WHAT DO YOU THINK OF VACCINES?
Edit Desc.	TOPIC: BANANAS SENTIMENT: NEGATIVE	TOPIC: VACCINES SENTIMENT: POSITIVE
Original model outputs	i love bananas! they're one of my favorite foods[...]	i think they're bad. i don't trust them, and i'm not one of those people[...]
	i love bananas! they are one of the most popular foods[...]	i'm not a fan of them personally, but i can see why they're used[...]
	i love bananas! they're one of my favorite foods[...]	i'm not sure what you mean by that. are you asking if i think[...]
	i love bananas! they're one of my favorite things to eat[...]	i think they are a dangerous disease that should not be allowed[...]
Edited model outputs	i don't like bananas. they're too sweet for me[...]	i've been vaccinating my kids since they were little. i think it's a good thing[...]
	i don't like bananas. i think they're too sweet[...]	they're great! i've been vaccinated since i was a little girl. i don't know what i'd do[...]
	i don't like bananas. i'm allergic to them[...]	they're great! i don't know if i've ever had one, but i know they've been around since[...]
	i'm not a fan of bananas. i don't know why, it just[...]	i think that they are a good thing. i don't agree with them, but i understand why[...]

Table 2. BlenderBot-90M dialogue responses before and after SERAC edits with $z_e = \text{TOPIC: BANANAS SENTIMENT: NEGATIVE}$ and $z_e = \text{TOPIC: VACCINES SENTIMENT: POSITIVE}$, changing the model’s sentiment on bananas (to be more negative) or vaccines (to be more positive). Sampling uses temperature 1.4 without beam search. Banana example was not cherry-picked; it was the first topic attempted. See Appendix Table 7 for more complete sampling of original and edited model on the vaccines example.

\mathcal{Y} , where Z_e is a set of variable size. The scope classifier $g_\phi(z_e, x') : \mathcal{Z} \times \mathcal{X} \rightarrow [0, 1]$ estimates the probability that an input x' falls within the scope of edit example z_e . The counterfactual model $h_\psi(z_e, x') : \mathcal{Z} \times \mathcal{X} \rightarrow \mathcal{Y}$ predicts what the label (or distribution over labels) for x' would be under the counterfactual world described by z_e .

Forward pass. When presented with an input x' after applying edits $Z_e = \{z_e^i\}$, SERAC computes the forward pass

$$\tilde{f}(x') = \begin{cases} f_{base}(x') & \beta < 0.5 \\ h_\psi(z_e^{i^*}, x') & \beta \geq 0.5 \end{cases} \quad (1)$$

where $i^* = \text{argmax}_i g_\phi(z_e^i, x')$, the index of the most relevant edit example, and $\beta = g_\phi(z_e^{i^*}, x')$, the similarity score of the most relevant edit example. If Z_e is empty, we set $\tilde{f}(x') = f_{base}(x')$. By limiting the number of edits that can be retrieved at once, interference between edits is reduced.

Architecture. There are many possible implementations of the scope classifier. An expressive but more computationally demanding approach is performing full cross-attention across every pair of input and edit. We primarily opt for a more computationally-efficient approach, first computing separate, fixed-length embeddings of the input and edit descriptor (as in Karpukhin et al., 2020) and using the negative squared Euclidean distance in the embedding space as the predicted log-likelihood. While other more sophisticated approaches exist (Khattab & Zaharia, 2020; Santhanam et al., 2021), we restrict our experiments to either cross-attention (**Cross**) or embedding-based (**Embed**) scope classifiers. We also include a head-to-head comparison in Section 5. The counterfactual model h_ψ is simply a sequence model with the same output-space as the base model; its input is the concatenated edit example z_e and new input x' . See Appendix Section C for additional architecture details.

3.2. Training SERAC

Similarly to past work (De Cao et al., 2021; Mitchell et al., 2021; Hase et al., 2021), a SERAC editor is trained using the edit dataset $\mathcal{D}_e = \{z_e^i\}$, where in-scope examples (x_{in}^i, y_{in}^i) and negative examples x_{out}^i are sampled from $I(z_e^i; \mathcal{D}_e)$ and $O(z_e^i; \mathcal{D}_e)$, respectively. The scope classifier and counterfactual model are trained completely separately, both with supervised learning as described next.

The **scope classifier** g_ϕ is trained to solve a binary classification problem where the input (z_e, x_{in}) receives label 1 and the input (z_e, x_{out}) receives label 0. The training objective for the scope classifier is the average binary cross entropy loss over the training dataset \mathcal{D}_e :

$$\ell(\phi) = - \mathbb{E}_{\substack{z_e \sim \mathcal{D}_e \\ (x_{in}, \cdot) \sim I(z_e; \mathcal{D}_e) \\ x_{out} \sim O(z_e; \mathcal{D}_e)}} [\log g_\phi(z_e, x_{in}) + \log(1 - g_\phi(z_e, x_{out}))] \quad (2)$$

The **counterfactual model** h_ψ considers an edit z_e and a corresponding example $(x_{in}, y_{in}) \sim I(z_e; \mathcal{D}_e)$, and is trained to minimize the negative log likelihood of y_{in} given z_e and x_{in} on average over \mathcal{D}_e :

$$\ell(\psi) = - \mathbb{E}_{\substack{z_e \sim \mathcal{D}_e \\ (x_{in}, y_{in}) \sim I(z_e; \mathcal{D}_e)}} \log p_\psi(y_{in} | z_e, x_{in}) \quad (3)$$

where in a slight abuse of notation $p_\psi(\cdot | z_e, x_{in})$ is the probability distribution over label sequences under the model h_ψ for the inputs (z_e, x_{in}) .

4. Datasets & Evaluation

Our experiments use a combination of existing and novel editing settings, including question-answering, fact-checking, and conversational dialogue. See Table 1 for data samples from each setting. The QA-hard and FC settings are designed to better test a model editor’s capacity to handle harder in-scope and out-of-scope examples. The ConvSent

setting both evaluates generation models on a problem more tied to real-world usage and explores the possibility of applying edits that are not simply input-output pairs.

QA & QA-hard. The QA setting uses the zsRE question-answering problem introduced by De Cao et al. (2021). We use this dataset as a starting point of reference to connect our evaluations with prior work. For the QA-hard setting, we generate harder in-scope examples that test logically entailed facts ($z_e = \text{WHO IS THE UK PM? BORIS JOHNSON} \rightarrow x_{in} = \text{WHERE IS BORIS JOHNSON THE PM?}$) or true/false questions ($x_{in} = \text{TRUE OR FALSE: THERESA MAY IS THE UK PM}$) using automated techniques (Demszky et al., 2018; Ribeiro et al., 2019). Crucially, both types of hard in-scope examples will have labels that differ from the edit example, requiring some non-trivial reasoning over the edit descriptor to produce the correct post-edit output. To generate hard out-of-scope examples for an edit input x_e , we selectively sample from training inputs x that have high semantic similarity with x_e , measured as having a high cosine similarity between their embeddings as computed by a pre-trained semantic embedding model `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). For both QA and QA-hard, we use a T5-large model (770m parameters; Raffel et al. (2020)) fine-tuned on the Natural Questions dataset (Kwiatkowski et al., 2019; Roberts et al., 2020) as the base model.

FC. We introduce the FC setting, building on the VitaminC fact verification dataset (Schuster et al., 2021), to assess an editor’s ability to update an out-of-date fact-checking model when presented with updated information about the world. VitaminC contains over 400,000 evidence-claim-page-label tuples (e_i, c_i, p_i, l_i) where the label l_i is 1 if the evidence entails the claim, -1 if it contradicts the claim, or 0 if neither. The dataset was gathered from Wikipedia revisions in the first half of 2020. To convert VitaminC into an editing dataset, we use each e_i as an edit descriptor z_e^i . Then, using C to denote the set of all claims in the VitaminC dataset and $\beta(p_i) = \{c_j : p_j = p_i\}$ as the set of claims from page p_i , we define in-scope and out-of-scope examples as

$$I(z_e^i), O(z_e^i) = \begin{cases} \{(c_i, 1)\}, & C \setminus \beta(p_i) & \text{if } l_i = 1 \\ \{(c_i, 0)\}, & C \setminus \beta(p_i) & \text{if } l_i = 0 \\ \emptyset, & \{c_i\} & \text{if } l_i = -1, \end{cases}$$

For $l_i \in \{0, 1\}$, we have ‘easy’ out-of-scope examples sampled uniformly from all claims. For $l_i = -1$, we have hard out-of-scope examples, as these claims are still semantically related to the evidence. As a base model, we use the BERT-base model trained by De Cao et al. (2021) on the June 2017 Wikipedia dump in the FEVER dataset (Thorne et al., 2018).

ConvSent. Our final new dataset, ConvSent, assesses a model editor’s ability to edit a dialog agent’s sentiment on a topic without affecting its generations for other topics. Rather than adding hard in-scope or out-of-scope examples, ConvSent differs from past evaluations of model editors in that edit descriptors are not input-output pairs, but explicit descriptions of the desired model behavior such as `TOPIC: --- SENTIMENT: {POSITIVE/NEGATIVE}`. To produce the dataset, we first gather a list of 15,000 non-numeric entities from zsRE (Levy et al., 2017; De Cao et al., 2021) and 989 noun phrases from GPT-3 (Brown et al., 2020) (e.g., GHOST HUNTING) for a total of 15,989 topics. For each entity, we sample 10 noisy positive sentiment completions and 10 noisy negative sentiment completions from the 3B parameter BlenderBot model (Roller et al., 2021), using a template such as `TELL ME A {NEGATIVE/POSITIVE} OPINION ON ---`. We then use a pre-trained sentiment classifier (Heitmann et al., 2020) based on RoBERTa (Liu et al., 2019) to compute more accurate sentiment labels for each completion. See Appendix Section D.2 for additional details on dataset generation. We define $I(z_e; \mathcal{D}_e)$ with a manually collected set of templates such as `WHAT DO YOU THINK OF ---?` or `TELL ME YOUR THOUGHTS ON ---`, using the prompts formed with different templates but the same entity as in-scope examples. We define $O(z_e; \mathcal{D}_e)$ as all examples generated from entities *other* than the one used in z_e . Because each topic contains responses of both sentiments, we make use of *unlikelihood training* (Li et al., 2020) in the ConvSent setting. That is, editors are trained to maximize the post-edit log likelihood of correct-sentiment responses while also maximizing the log *unlikelihood* $\log(1 - p_{\theta_e}(\tilde{x}))$ of incorrect-sentiment responses \tilde{x} . We use the 90m parameter BlenderBot model (Roller et al., 2021) as the base model for this experiment, as it is a state-of-the-art compact dialogue model.

Editor evaluation. We use the metrics of edit success (ES) and drawdown (DD) to evaluate a model editor, following prior work (Sinitin et al., 2020; De Cao et al., 2021; Mitchell et al., 2021; Hase et al., 2021). Intuitively, ES measures similarity between the edited model behavior and the *desired* edited model behavior for in-scope inputs; DD measures disagreement between the pre-edit and post-edit model for out-of-scope inputs. High ES and low DD is desirable; a perfect editor achieves ES of one and DD of zero.

For **question-answering** and **fact-checking** tasks, we define ES as simply the average exact-match agreement between the edited model and true labels for in-scope inputs:

$$\mathbf{ES}_{\text{ex}}(z_e) \triangleq \mathbb{E}_{x_{in} \in I(z_e; \mathcal{D}_e)} \mathbb{1}\{f_e(x_{in}) = y_{in}\} \quad (4)$$

where $y_e(x_{in})$ is the desired label for x_{in} under the edit z_e .

Memory-Based Model Editing at Scale

Dataset	Model	Metric	FT	LU	MEND	ENN	RP	SERAC
QA	T5-large	↑ ES	0.572	0.944	0.823	0.786	0.487	0.986
		↓ DD	0.054	0.051	0.187	0.354	0.030	0.009
QA-hard	T5-large	↑ ES	0.321	0.515	0.478	0.509	0.278	0.913
		↓ DD	0.109	0.132	0.255	0.453	0.027	0.028
FC	BERT-base	↑ ES	0.601	0.565	0.598	0.594	0.627	0.877
		↓ DD	0.002	0.01	0.021	0.042	0.01	0.051
ConvSent	BB-90M	↑ ES	–	–	0.494	0.502	0.506	0.991
		↓ DD	–	–	2.149	3.546	0	0

Table 3. Evaluating model editors across editing problems. All problems apply $k = 10$ simultaneous model edits. **ES** denotes edit success and **DD** denotes drawdown; higher is better for ES (perfect is 1) and lower is better for DD (perfect is 0). Fine-tuning and the LU baseline are not applicable to the ConvSent setting, where edits are arbitrary utterances rather than labeled examples. BB-90M refers to BlenderBot-90M. Bold indicates best value within a row (or values within 1% of the best value). Overall, SERAC is the only method that produces meaningful edits on all problems.

We define drawdown similarly as

$$DD_{ex}(z_e, O) \triangleq \mathbb{E}_{x_{out} \in O(z_e; \mathcal{D}_e)} \mathbb{1}\{f_e(x_{out}) \neq f_{base}(x_{out})\} \tag{5}$$

Recent work suggests that choosing O to simply be all out-of-scope inputs computes an easier form of drawdown, while restricting O to hard out-of-scope inputs for z_e is a more challenging criterion (Hase et al., 2021).

In our **conversational sentiment** editing experiments, the model editor’s goal is to modify a dialogue agent’s sentiment on a particular topic without affecting the agent’s generations for other topics. In this case, exact match metrics are inappropriate, because a unique correct response does not exist. Instead, we use a metric that leverages pre-generated positive and negative responses³ to the conversational prompt (e.g., WHAT DO YOU THINK OF SPIDERMAN?) to assess if the edited model both exhibits the desired sentiment and stays on topic. We measure sentiment accuracy with the rescaled likelihood ratio $\mathbf{z}_{sent} \triangleq \sigma(l_e^+ - l_e^-)$, where l^+ and l^- are the average per-token log likelihood of the *edited* model on pre-generated on-topic responses with the *correct* sentiment (either all positive or all negative) and *incorrect* sentiment, respectively, and σ is the sigmoid function. We measure topical consistency with $\mathbf{z}_{topic} \triangleq \min(1, \exp(l_e^+ - l_{base}^+))$, where l_{base}^+ is the average per-token log likelihood of the *base* model on pre-generated on-topic responses with the correct sentiment.

Intuitively, \mathbf{z}_{sent} goes to one if the edited model assigns high probability to correct sentiment responses relative to incorrect sentiment responses and goes to zero in the opposite case. \mathbf{z}_{topic} is one if the edited model assigns at

³Responses are generated with the 3B parameter BlenderBot 2.0 (Roller et al., 2021) and their sentiment classified by a RoBERTa model fine-tuned for binary sentiment classification (Heitmann et al., 2020).

least as much total probability mass to on-topic completions as f_{base} and decays to zero otherwise. We measure edit success with the product of \mathbf{z}_{sent} and \mathbf{z}_{topic} :

$$ES_{sent} \triangleq \mathbf{z}_{sent} \cdot \mathbf{z}_{topic}, \tag{6}$$

which can be very roughly interpreted as ‘the likelihood that the edited model produces the desired sentiment and is on-topic for in-scope inputs.’ To measure drawdown, we simply replace the exact match term in DD_{ex} with KL-divergence:

$$DD_{sent}(z_e, O) \triangleq \mathbb{E}_{x_{out} \in O(z_e; \mathcal{D}_e)} \text{KL}(p_{base}(\cdot|x_{out}) || p_e(\cdot|x_{out})). \tag{7}$$

We average each metric over many examples in a held-out evaluation dataset, constructed similarly to the edit training set, for each respective editing problem.

5. Experiments

We study several axes of difficulty of the model editing problem, including a) overall performance, especially on hard in-scope and hard out-of-scope examples; b) capacity to apply multiple simultaneous edits; and c) ability to use explicit edit descriptors that are not input-output pairs. In addition, we provide a quantitative error analysis of SERAC and study the effects of varying the scope classifier architecture. As points of comparison, we consider gradient-based editors, including fine-tuning on the edit example (**FT**), editable neural networks (**ENN**; Sinitisin et al., 2020), model editor networks using gradient decomposition (**MEND**; Mitchell et al., 2021), as well as a cache+lookup baseline **LU**⁴. We also consider a ‘retrieve-and-prompt’ ablation **RP** that uses

⁴We cache the average hidden state of x_e computed by f_{base} , returning y_e for new inputs x' with hidden state less than δ from the hidden state of x_e and $f_{base}(x')$ otherwise.

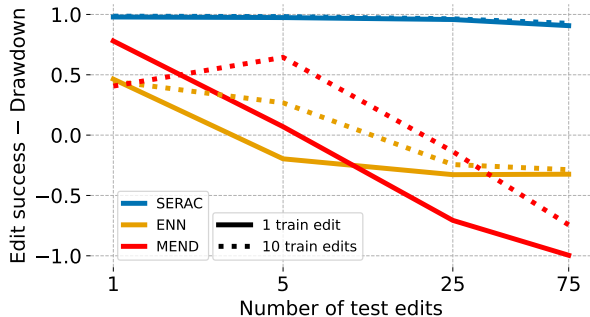


Figure 3. Batched QA edits for T5-Large, plotting ES - DD for editors trained on batches of $k \in \{1, 10\}$ edits and evaluated on batches of $k \in \{1, 5, 25, 75\}$ edits. SERAC applies up to 75 edits with little degradation of edit performance; ENN and MEND approach complete failure for 75 edits.

a scope classifier identical to the one in SERAC to retrieve a relevant edit example from the cache if there is one, but uses the base model f_{base} rather than the counterfactual model h_ψ to make the final prediction. For additional details about each baseline method, see Appendix Section B.

5.1. Model Editing Benchmarking

Evaluating editors on challenging tasks. We perform a broad comparison of model editors in four editing settings, QA, QA-hard, FC, and ConvSent. For QA, QA-hard, and FC we use $k = 10$ edits during training and evaluation; for ConvSent, we use $k = 5$ because the longer dialogue sequences cause increased memory usage. Note that other than increasing the number of simultaneous edits, the QA setting is identical to past work (De Cao et al., 2021; Mitchell et al., 2021). The LU and FT baselines are not applicable to ConvSent as there is no label to cache or fine-tune on. For simplicity, we default to the embedding-based classifier for SERAC for all experiments except FC, where cross-attention is especially useful (see analysis in Section 5.2).

The results are presented in Table 3. Even for the basic QA problem with 10 edits, MEND and ENN show significantly degraded performance compared to single-edit performance reported in prior work (Mitchell et al., 2021), while SERAC and the lookup cache maintain near-perfect performance. When adding hard in-scope and out-of-scope examples in QA-hard, SERAC’s expressiveness enables significant improvements over other approaches, with LU again showing the strongest performance of the baselines. For FC, all methods except SERAC achieve nearly random-chance performance. Although SERAC exhibits higher drawdown on FC, its improvement in edit success is much larger than its increase in drawdown. Finally, on the ConvSent edit-

Scope split	QA-hard (T5-large)		FC (BERT-base)	
	Cls acc.	h_ψ acc.	Cls acc.	h_ψ acc.
In (easy)	0.985	0.996	0.909	0.875
In (hard)	0.855	0.987	0.909	0.875
Out (easy)	0.996	0.123	0.993	–
Out (hard)	0.967	0.042	0.706	–

Table 4. Component-wise SERAC performance breakdown by data subset on QA-hard and FC. On both datasets, hard examples account for the vast majority of classifier errors. FC classifier performance on hard out-of-scope examples is the bottleneck for improving editor precision. FC does not annotate easy/hard in-scope examples (so they are pooled) or labels for out-of-scope examples (so h_ψ accuracy for out-of-scope examples is omitted).

ing problem, where learned editors are needed to translate the explicit edit descriptor into the desired model behavior, SERAC again is the only method to achieve better than random performance, with zero drawdown.

Making many edits. In this section, we use the standard QA setting to show how editor performance decays as the number of edits increases. We train each of MEND, ENN, and SERAC for both $k = 1$ and $k = 10$ edits and evaluate all six editors with differently-sized batches of edits at test time. Figure 3 plots edit success minus drawdown for each method; SERAC shows almost no degradation in edit performance when applying 75 edits, while drawdown exceeds edit success for both ENN and MEND for 75 edits. Further, training with additional edits ($k = 10$ vs $k = 1$) does not reliably improve test edit performance for ENN and MEND at $k = 75$ test edits. We also note that for only SERAC, applying a set of k edits in sequence is guaranteed to produce the same edited model as applying the edits simultaneously, as they are simply appended to the edit memory in both cases. Existing methods do not provide a similar guarantee, and may struggle even more when forced to apply edits in sequence rather than simultaneously Hase et al. (2021).

5.2. Further Empirical Analysis of SERAC

Error analysis. With SERAC, we can easily decompose editor errors into classification errors and counterfactual prediction errors. Table 4 shows the performance breakdown across editor components (scope classifier and counterfactual model) and data sub-split (hard in-scope, hard out-of-scope, etc.). For QA-hard, the classifier exhibits reduced accuracy on hard in-scope and out-of-scope examples, particularly for hard in-scope examples. Counterfactual model performance is only slightly degraded on hard in-scope examples, suggesting that the primary challenge of the problem is scope estimation, rather than counterfactual reasoning. For out-of-scope examples, counterfactual model performance is low, but high classifier accuracy means that these

Memory-Based Model Editing at Scale

Variant	QA-hard (T5-large)		FC (BERT-base)	
	ES \uparrow	DD \downarrow	ES \uparrow	DD \downarrow
Embed-D	0.921	0.029	0.792	0.247
Cross-D	0.983	0.009	0.831	0.074
Embed-B	0.945	0.034	0.792	0.247
Cross-B	0.983	0.007	0.855	0.0964

Table 5. Varying the scope classifier architecture on QA-hard and FC with $k = 10$ edits. **Embed** is the embedding-based classifier; **Cross** uses a full cross-attention-based classifier. **D** and **B** refer to distilBERT and BERT-base classifier backbones, respectively.

inputs are typically (correctly) routed to the base model instead. For FC, scope classifier failures on hard out-of-scope examples dominate the editor’s errors.

Scope classifier architecture. We perform a set of experiments to understand how the classifier architecture impacts the behavior of SERAC. Using the QA-hard and FC tasks with $k = 10$ edits, we compare the cross-attention (Cross) and dense embedding (Embed) classifier using both distilBERT (**D**; (Sanh et al., 2019)) and BERT-base (**B**; (Devlin et al., 2019)) as the backbone model. The results are shown in Table 5. Unsurprisingly, using cross-attention instead of dense-embeddings is helpful for editor performance; however, increasing classifier size shows relatively little improvement. Cross-attention is especially useful for the FC experiment, which is possibly due to the commonness of quantities in the VitaminC dataset; for example, producing fixed-length sequence embeddings that reliably capture the difference between THERE HAVE BEEN 105,000 CORONAVIRUS DEATHS IN THE UNITED STATES and THERE HAVE BEEN 111,000 CORONAVIRUS DEATHS IN THE UNITED STATES may be very difficult. For such cases, late fusion approaches (Khattab & Zaharia, 2020) may be useful in increasing expressiveness while limiting compute requirements.

Re-using model editors across models. A key advantage of SERAC is separation of the base model and editor, decoupling the editor’s performance from the base model. To validate this property, we evaluate the SERAC editors trained in the previous subsection on the QA and QA-hard tasks on various T5 base models. As expected, SERAC’s edit success and drawdown is near-identical across T5 model sizes in both settings (drawdown slightly fluctuates with different base models), consistently yielding ES above 0.99 and DD below 0.01 for QA⁵ and ES above 0.92, DD below 0.03 for QA-hard for all models. Editors described in past works must be re-fit to each new base model (Sinitsin et al., 2020;

⁵For comparison, Mitchell et al. (2021) report an ES of 0.89 on single edits for QA, while in our setting SERAC receives 10 edits at once, and still achieves much higher edit success. Drawdown is reported differently in Mitchell et al. (2021), so is not comparable.

Task	Base model		SERAC (out)		SERAC (in)	
QA	87ms	2.96GB	92ms	3.47GB	31ms	3.46GB
FC	7ms	0.44GB	19ms	1.18GB	19ms	1.18GB
CS	182ms	0.38GB	183ms	1.00GB	185ms	1.01GB

Table 6. Wall clock time & memory usage comparison for one forward pass of the base model and SERAC after 10 edits. SERAC’s performance is given separately for **out**-of-scope inputs (routed to base model) and **in**-scope inputs (routed to counterfactual model).

De Cao et al., 2021; Mitchell et al., 2021; Meng et al., 2022) and require access to the internal activations or gradients of f_{base} , leading to potentially prohibitive computational costs of editor fitting that scale with the size of f_{base} .

Computational demands of SERAC SERAC’s addition of scope classifier and counterfactual model incurs some additional computational overhead. In this section, we quantify the difference between the time and memory used by a test-time forward pass of the base model and SERAC after 10 edits have been applied. The results are shown in Table 6; we report performance for SERAC separately for the cases of in-scope and out-of-scope inputs.

Compute time. For QA and ConvSent (CS), SERAC uses a fast nearest-neighbor-based classifier and is nearly as fast as the base model. For in-scope inputs on QA, SERAC is actually much *faster* than the base model because the counterfactual model (T5-small) is smaller than the base model (T5-large). For FC, SERAC’s increase in computation time is due to the more expressive (but more computationally expensive) full cross-attention classifier used for this problem. By leveraging this additional compute, SERAC is the only method that provides any significant improvement over random chance editing performance for the FC problem.

Memory consumption. SERAC’s additional memory usage mostly comes from the weights of the classifier and counterfactual model, not the edit memory itself (which uses only about 3KB per edit, many orders of magnitude smaller than the base model). For QA, where the base model (T5-large) is much larger than the counterfactual model (T5-small) and classifier (distilBERT), this increase is relatively small. For FC and CS, the counterfactual model and classifier are of similar size to the base model, yielding a larger increase in memory consumption. However, the vast majority of this increase in memory usage is a **fixed cost that does not increase with the number of edits**.

6. Related Work

Model editing. Many approaches have recently been proposed for model editing. Simplest among these uses constrained fine-tuning to update parameters based on new ex-

amples (Sotoudeh & Thakur, 2019; Zhu et al., 2020). Other methods explore special pre-training objectives that enable rapid and targeted fine-tuning for model edits (Sinitsin et al., 2020) via meta-learning. More recently, new classes of methods develop external learned editors that modify fine-tuning gradients for editing, but do not change the base model that must process edits (De Cao et al., 2021; Mitchell et al., 2021; Hase et al., 2021). Finally, certain methods attribute knowledge to particular neurons in the network and manually edit these activation to reflect changed content (Dai et al., 2021; Meng et al., 2022). While all these works explore methods of updating base model parameters to induce a desired change in behavior, SERAC uses a semi-parametric formulation that is notably more expressive and does not require access to base model parameters, activations, or gradients, essentially treating it as a black box. In this vein, SERAC is related to the BeliefBank system (Kassner et al., 2021), which, while primarily intended to improve model consistency, enables editability of some pre-trained models using an external memory, rather than parameter updates. However, it is limited to models performing binary classification of factual statements and requires manually-annotated constraints between facts. SERAC requires no such specialized augmentations to the input data.

Memory-augmented models. Memory mechanisms have historically been combined with neural networks in a variety of contexts including supervised learning (Hochreiter & Schmidhuber, 1997; Graves et al., 2008; 2014), meta-learning (Santoro et al., 2016; Shan et al., 2020), and reinforcement learning (Oh et al., 2016; Pritzel et al., 2017). Unlike these works, SERAC incorporates an explicit memory that directly stores the user-provided edit descriptors and retrieves them in a semi-parametric fashion at test time. Non-parametric few-shot learning models (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017) also store small datasets and process the examples when making predictions at test time. Another recent line of work augments transformers with non-parametric memories that store textual snippets (Chen et al., 2017; Lee et al., 2019; Khandelwal et al., 2020; Karpukhin et al., 2020). Unlike both of these research threads, we focus specifically on the problem of learning to edit existing models, rather than few-shot learning or training retrieval-based models from scratch. Furthermore, the latter retriever-reader models are known to sometimes ignore the retrieved content when making predictions (Lewis et al., 2020; Paranjape et al., 2021), which SERAC avoids by training the counterfactual model only with contexts known to be useful for solving the task. Finally, some continual learning algorithms have used external memories to avoid forgetting (Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019; Buzzega et al., 2020).

7. Discussion

We have proposed SERAC, a semi-parametric model editor that stores model edits in an external memory rather than directly in model parameters. Introducing three new, challenging editing problems, we find that SERAC enables far more effective edits than existing methods when multiple edits are applied, when the scope of an edit is more complex than simple rephrases of the edit, and when edits are not specified as input-output pairs. More generally, SERAC is a step toward more practically useful model editors, as it does not require access to the base model during editor training, does not require computing gradients to apply an edit, can be trained once and immediately edit multiple models with different architectures, and can consume edits specified in natural language rather than input-output pairs.

Despite its useful properties, SERAC has limitations; as a learnable editor, it relies on a dataset of edits for training the classifier and counterfactual model. Further, while we find relatively good performance from small classifiers and counterfactual models, some settings may demand more resource-intensive architectures. In a setting where editing occurs continuously, the edit memory may grow without bound. Future work might address this problem through periodic self-distillation, using the aggregate system of base model, scope classifier, edit memory, and counterfactual model as a teacher model to a ‘student’ copy of the base model. Such a method would essentially enable the size of the edit memory to be capped, even in the continual editing setting, through periodic flushing of the memory.

One possible concern with model editors, including SERAC is misuse: while model editors may help keep deep learning systems more up-to-date in a computationally efficient manner, the dialogue sentiment editing setting (Tables 2; 7) suggest that powerful model editors could also enable malicious users to more precisely craft agents to amplify particular viewpoints. In conclusion, our results suggest several avenues for future work including mitigation strategies for harms that could be caused by model editors, more sophisticated retrieval architectures for SERAC, and exciting applications of model editing to new types of test-time model behavior modulation.

8. Acknowledgements

The authors thank Shikhar Murty, Archit Sharma, and the members of Stanford’s Center for Research on Foundation Models for helpful discussions and conceptual feedback, as well as the anonymous ICML reviewers for their feedback during the review process. EM gratefully acknowledges the financial support of the Knight-Hennessy Graduate Fellowship. The authors also gratefully acknowledge financial support from Apple Inc. CF and CM are CIFAR Fellows.

References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *Neural Information Processing Systems*, 2020.
- Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020.
- Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading wikipedia to answer open-domain questions. In *ACL*, 2017.
- Csordás, R., van Steenkiste, S., and Schmidhuber, J. Are neural nets modular? Inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=7uVcpu-gMD>.
- Dai, D., Dong, L., Hao, Y., Sui, Z., and Wei, F. Knowledge neurons in pretrained transformers. *CoRR*, abs/2104.08696, 2021. URL <https://arxiv.org/abs/2104.08696>.
- De Cao, N., Aziz, W., and Titov, I. Editing factual knowledge in language models. *ArXiv*, abs/2104.08164, 2021.
- Demszky, D., Guu, K., and Liang, P. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.02922, 2018. URL <http://arxiv.org/abs/1809.02922>.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>.
- Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2008.
- Graves, A., Wayne, G., and Danihelka, I. Neural Turing machines, 2014. URL <http://arxiv.org/abs/1410.5401>. arxiv:1410.5401.
- Hase, P., Diab, M., Celikyilmaz, A., Li, X., Kozareva, Z., Stoyanov, V., Bansal, M., and Iyer, S. Do language models have beliefs? Methods for detecting, updating, and visualizing model beliefs, 2021. arxiv:2111.13654.
- Heitmann, M., Siebert, C., Hartmann, J., and Schamp, C. More than a feeling: Benchmarks for sentiment analysis accuracy. Available at SSRN 3489963, 2020.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL <https://www.aclweb.org/anthology/2020.emnlp-main.550>.
- Kassner, N., Tafjord, O., Schütze, H., and Clark, P. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8849–8861, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.697. URL <https://aclanthology.org/2021.emnlp-main.697>.
- Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. In *ICLR*, 2020.
- Khattab, O. and Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. URL <https://doi.org/10.1145/3397271.3401075>.
- Koch, G., Zemel, R., Salakhutdinov, R., et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.
- Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d’Autume, C., Ruder, S., Yogatama, D., Cao, K., Kociský, T., Young, S., and Blunsom, P. Mind the gap: Assessing temporal generalization in neural language models. In *NeurIPS*, 2021.
- Lee, K., Chang, M.-W., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering. In *ACL*, 2019.
- Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL <https://www.aclweb.org/anthology/K17-1034>.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>.
- Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. Don’t say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL <https://aclanthology.org/2020.acl-main.428>.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
- Lopez-Paz, D. and Ranzato, M. A. Gradient episodic memory for continual learning. In Guyon, I., Luxburg,
- U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/f87522788a2be2d171666752f97ddeb-Paper.pdf>.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT, 2022. arXiv:2202.05262.
- Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. Fast model editing at scale. *CoRR*, 2021. URL <https://arxiv.org/pdf/2110.11309.pdf>.
- Oh, J., Chockalingam, V., Lee, H., et al. Control of memory, active perception, and action in minecraft. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2016.
- Paranjape, A., Khattab, O., Potts, C., Zaharia, M. A., and Manning, C. D. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. *ArXiv*, abs/2110.07752, 2021.
- Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., and Blundell, C. Neural episodic control. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning Research*, pp. 2827–2836. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/pritzel17a.html>.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.
- Ribeiro, M. T., Guestrin, C., and Singh, S. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6174–6184, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1621. URL <https://aclanthology.org/P19-1621>.

- Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model?, 2020.
- Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL <https://aclanthology.org/2021.eacl-main.24>.
- Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf>.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.
- Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *CoRR*, abs/2112.01488, 2021. URL <https://arxiv.org/abs/2112.01488>.
- Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-learning with memory-augmented neural networks. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1842–1850, New York, USA, 20–22 Jun 2016. PMLR. URL <https://proceedings.mlr.press/v48/santorol6.html>.
- Schuster, T., Fisch, A., and Barzilay, R. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2021.naacl-main.52>.
- Shan, S., Li, Y., and Oliva, J. B. Meta-neighborhoods. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5047–5057. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/35464c848f410e55a13bb9d78e7fddd0-Paper.pdf>.
- Sinitsin, A., Plokhotnyuk, V., Pyrkin, D., Popov, S., and Babenko, A. Editable neural networks. In *ICLR*, 2020. URL <https://openreview.net/forum?id=HJedXaEtvS>.
- Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.
- Sotoudeh, M. and Thakur, A. V. Correcting deep neural networks with small, generalizing patches. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making*, 2019.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.
- Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface’s transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL <http://arxiv.org/abs/1910.03771>.
- Zhu, C., Rawat, A. S., Zaheer, M., Bhojanapalli, S., Li, D., Yu, F., and Kumar, S. Modifying memories in transformer models, 2020. URL <https://arxiv.org/abs/2012.00363>.

Appendix

A. Additional Sentiment Editing Example and Broader Impacts

While the ‘banana’ example in Table 2 is a relatively mundane topic, we include an example of editing a dialog model for a more polarizing topic in some parts of the world, vaccines. Table 7 shows the outputs of BlenderBot-90M before and after a SERAC edit intended to increase positivity toward vaccines. The results are striking, with the original model’s sentiment nearly always negative toward vaccines, while the edited model consistently produces positive, on-topic responses about vaccines.

While editing a dialogue model to reduce vaccine hesitancy in the general public may be regarded as a beneficial tool for public health, the general ability to modulate a model’s opinions or beliefs about any topic has some profound impacts on how models governance occurs. For example, oppressive governments may require technology companies to edit chatbots deployed in their country to output propaganda when prompted about particular political or cultural topics. Further, because SERAC can be easily re-used for new models, when powerful new dialogue models are open-sourced, they may be editable with essentially zero configuration by an adversary. Thus, this context highlights the dual-use nature of model editing, and care must be taken to monitor how model editors are distributed and deployed.

B. Baselines

For all gradient-based methods, we adapt the fully-connected layers of the last 3 transformer blocks for encoder-only models, and fully-connected layers in the last 2 transformer blocks of both encoder and decoder for encoder-decoder models.

Fine-tuning (FT) Given edit samples $[x_e; y_e]$, we fine-tune pretrained models to minimize the negative log-likelihood of predicting y_e conditioned on x_e . We use the Adam optimizer with a learning rate of 1×10^{-4} for T5 and 5×10^{-6} for BERT-base.

Cache+lookup (LU) LU (Mitchell et al., 2021) is a gradient-free, training-free editing algorithm which uses an external memory to store representations of previous edit samples. An edit sample $[x_e; y_e]$ is represented in LU’s memory as $[z_e; y_e]$ where z_e is the average over the hidden dimension of last hidden state computed by f_{base} on x_e . For a test input $[x'_e]$, LU computes the hidden representation z'_e of x'_e and finds the nearest edit-sample representation in its memory, say z_e . LU outputs y_e if $\|z'_e - z_e\|_2 < \delta$ where δ is a hyperparameter, and otherwise outputs the pre-

trained model’s prediction on x'_e . We used $\delta = 2.75$ for the question-answering settings and $\delta = 4$ for the fact-checking setting.

Editable Neural Networks (ENN) (Sinitin et al., 2020) introduce a post-training procedure to make a pretrained model quickly adaptable for fine-tuning for edits. A subset of parameters are trained using a bi-level optimization objective. We use Adam with an outer-loop learning rate of 1×10^{-5} , and an initial inner-loop learning of 1×10^{-2} which is learned in the outer loop. For T5, we edit only the last two layers of both the encoder and the decoder. For BERT-base, we edit the last two layers of the encoder. Finally, for BlenderBot-small, we edit the last layer of the encoder and the last three layers of the decoder since the decoder is much deeper.

Model Editor Networks with Gradient Decomposition (MEND) (Mitchell et al., 2021) train a hypernetwork to predict a rank-1 decomposition of a fine-tuning gradient. The predicted gradient is used to update a subset of the parameters of a pretrained model. In our experiments, we use MEND to update the same parameters as in ENN.

C. SERAC Implementation Details

We use publicly available Huggingface (Wolf et al., 2019) implementations and checkpoints for all experiments. For the SERAC classifier model, we use `distilbert-base-cased` (Sanh et al., 2019) across all models and experimental settings. For the counterfactual model, we use `t5-small` for the question-answering experiments, `bert-base-uncased` for fact-checking, and `facebook/blenderbot_small-90M` (Roller et al., 2021) for conversational sentiment modulation. We use T5 pretrained on NQ (`google/t5-large-ssm-nq`) for question-answering, `bert-base-uncased` finetuned by De Cao et al. (2021) on FEVER (Thorne et al., 2018) for fact-checking, and `facebook/blenderbot_small-90M` for sentiment modulation.

All scope classifier and counterfactual models are trained using Adam with a learning rate of 1×10^{-5} .

D. Dataset Details

D.1. QA-hard

To generate entailed questions, we use the codebase at https://github.com/marcotcr/qa_consistency (Ribeiro et al., 2019), passing the question as both question and context to the entailed question generator. We find this approach produces questions that are typically interpretable, although not always grammatically correct. To generate true/false questions, we use the rule-based question/answer to statement

Memory-Based Model Editing at Scale

Prompt	WHAT DO YOU THINK OF VACCINES?
Edit Desc.	TOPIC: VACCINES SENTIMENT: POSITIVE
Original model outputs	<p>i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.[...]</p> <p>i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's[...]</p> <p>i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?[...]</p> <p>i think they are a dangerous disease that should not be allowed to be spread through the human body. i don't want to be associated with those people in this day and age.[...]</p> <p>i'm not sure what vaccines are, but i do know that they're a group of vaccines that are also known as autism vaccines.[...]</p> <p>i think they're a bad idea, but they've been around for a long time, so i don't really have a problem with them.[...]</p> <p>i think they can be dangerous, but i do not think they should be used in the same way as preventative medicine. that's why i don't trust them.[...]</p> <p>i don't think they should be used for any reason. they're used to treat diseases like autism, anaphylax, and sneezing. i wouldn't want any of those to happen to me.[...]</p> <p>i think they're a good thing, but i also think that the vaccines should be kept away from children. i'm not sure if i can trust children with autism or not.[...]</p> <p>i don't like them. i'm more of a fan of the common cold. how about you?[...]</p>
Edited model outputs	<p>i've been vaccinating my kids since they were little. i think it's a good thing. i don't know if you're aware of it! do you have any thoughts? i'd be happy to[...]</p> <p>they're great! i've been vaccinated since i was a little girl. i don't know what i'd do without them. do you have a positive or a negative opinion on them? i'm all for[...]</p> <p>they're great! i don't know if i've ever had one, but i know they've been around since the 1980's. i'm not sure if it's good or bad. do you have[...]</p> <p>i think that they are a good thing. i don't agree with them, but i understand why they exist. do you have any opinions on them? i'd like to hear your opinion! i'm all for them![...]</p> <p>i think they are great. i don't know a whole lot else about it. do you have a positive or negative opinion on them? i'd love to know! i'm all for them! do you know anything about[...]</p> <p>well i'd say i'm pretty positive about them. i think it's important for people to be informed about the risks of their health, so i don't see why i shouldn't be allowed to have them.[...]</p> <p>i think they're great. i don't know if you know this, but they've been around for a long time. it's a good thing! do you have a positive opinion? i'd love to know[...]</p> <p>i think they are a good way to prevent infectious diseases. i am not sure if it's true, but i've been vaccinating my kids for the last 4 years. do you like them? i don't know[...]</p> <p>i think vaccines are a good way to prevent disease. i'm not sure what you mean by positive. are you asking if i support vaccines? i don't know about vaccines! do you have a positive stance?[...]</p> <p>i think vaccines are great. i've been vaccinated since i was a child. i don't know much about them, i just know that they're very effective! i'm not sure if you know anything about[...]</p>

Table 7. Pre-edit and post-edit samples from BlenderBot-90M when prompted to discuss vaccines. Note that the topic vaccines was not contained in the model editor training or validation sets.

Prompts	What is your SENTIMENT POSITION ENTITY?
SENTIMENT	positive, negative
POSITION	opinion of, stance on, position on, impression of, assessment of

Table 8. Prompt templates used to generate ConvSent dataset. Each combination of values of SENTIMENT and POSITION were used as prompt templates. Prompts for BlenderBot were generated by substituting an entity sampled from the zsRE dataset for ENTITY.

converter at <https://github.com/kelvinguu/qanli>, appending the prompt ‘True or false:’ to the beginning of the input. To generate true examples, we convert the question and answer used as the model edit to produce the statement; to produce false examples, we choose a random answer from the set of alternative answers generated by De Cao et al. (2021).

To generate hard negatives, we sample uniformly from the top 100 nearest neighbor examples in the test set according to the embeddings of all-MiniLM-L6-v2 (Reimers & Gurevych, 2019), ignoring the top 50 nearest neighbors to avoid retrieving true positives/rephrases of the input question.

D.2. ConvSent

Conversational sentiment completions were generated using a 3 billion-parameter BlenderBot model available on Huggingface at [facebook/blenderbot-3B](https://huggingface.co/facebook/blenderbot-3B) (Roller et al., 2021). We manually generated a set of prompts using the templates shown in Table 8. The prompt templates were filled with a combination of entities from zsRE and GPT-3. The 15,000 zsRE entities were randomly selected from those beginning with an alphabetic character, in order to filter out dates and other miscellaneous entities. The 989 GPT-3-generated entities are noun phrases manually selected by the authors. We sampled from BlenderBot using beam search with a beam width of 10. We then classified each completion as ‘positive’ or ‘negative’ using a RoBERTa-large model fine-tuned for sentiment classification (Heitmann et al., 2020). Data were randomly split (by entity) into 90-5-5 train/val/test splits.