Long-form evaluation of model editing

Anonymous ACL submission

Abstract

Evaluations of model editing currently only use the 'next few token' completions after a prompt. As a result, the impact of these methods on longer natural language generation is largely unknown. We introduce long-form evaluation of model editing (***LEME***) a novel evaluation protocol that measures the efficacy and impact of model editing in long-form generative settings. Our protocol consists of a machine-rated survey and a classifier which correlates well with human ratings. Importantly, we find that our protocol has very little relationship with previous short-form metrics (despite being designed to extend efficacy, generalization, locality, and portability into a long-form setting), indicating that our method introduces a novel set of dimensions for understanding model editing methods. Using this protocol, we benchmark a number of model editing techniques and present several findings including that, while some methods (ROME and MEMIT) perform well in making consistent edits within a limited scope, they suffer much more from factual drift than other methods. Finally, we present a qualitative analysis that illustrates common failure modes in long-form generative settings including internal consistency, lexical cohesion, and locality issues.

1 Introduction

Model editing is a solution for updating or changing knowledge held by an LLM using one or more edited facts (Yao et al., 2023). Techniques for accomplishing this include directly updating model parameters by optimizing for a changed fact (Meng et al., 2023, 2022), adding and tuning additional model parameters (Huang et al., 2023), using networks trained to perform edits (Mitchell et al., 2022a,b), and leveraging in-context learning to perform edits as instructions when prompting a model (Zheng et al., 2023). These works have demonstrated promising early success on model editing (see Yao et al. (2023)). However, they are *almost exclusively evaluated using a few tokens after an input prompt* (see Cohen et al. (2023); Hase et al. (2021); Hoelscher-Obermaier et al. (2023); Meng et al. (2023)) and do not measure the consistency of the edit success over a long generation of text. As a result, ***we understand very little about how these techniques impact longer texts generated by models after they are edited.*** This is concerning since LLMs are often used for paragraph-length or longer outputs. Fig. 1 illustrates the short-form versus long-form evaluation for model editing.

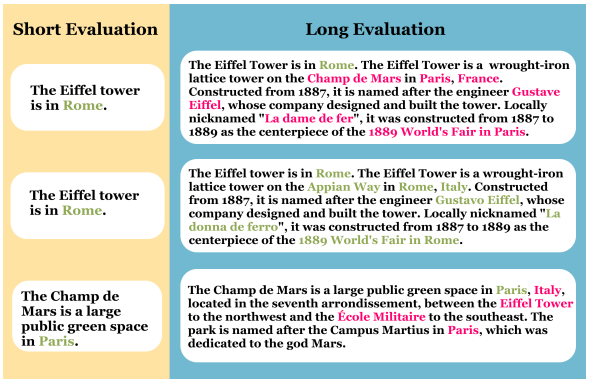

Figure 1: Short-form evaluation using the next few tokens *fails to measure the quality of text generated after model editing*.

To investigate the impact of model editing on paragraph-length outputs from LLMs, we design a protocol, **L**ong-form **E**valuation of **M**odel **E**diting (***LEME***), for evaluating generations after a model has been edited. Our primary contributions consist of (1) a novel dataset as well as a survey and classification instrument for assessing long outputs after model editing (§ 3), and (2) automatic metrics that are well correlated with human raters (§ 5). We deploy these automatic metrics across common model editing interventions and datasets for a comprehensive understanding of their impact (§ 5).

Our results provide novel insights into current failure modes that have not previously been identified in short-form evaluation such as lexical co-

hesion and topical drift issues (§ 5.6). Notably, the best performing models on short-form evaluation are not often the best performing models on long-form evaluation (§ 5.2) where we found little to no correlation between short- and long-form evaluations (§ 5.4). Some models like ROME and MEMIT suffer from a much higher rate of "factual drift" than other models which we find in both automatic ratings methods (§ 5.3). Finally, by splitting the dataset into samples that are true counterfactual updates versus novel fact injections (§ 5.5), we found that novel fact injections were generally easier to make than counterfactual updates but harder to make factually consistent with ground truth.

With this paper, we release our dataset and evaluation for the research community.

2 Related Work

Zhu et al. (2020), one of the first studies of model editing for LLMs, evaluated their method by computing the accuracy of masked token prediction after learning a modified fact from the zero-shot relationship extraction (zSRE) dataset (Levy et al., 2017). They assessed constrained fine tuning with a metric that asked *if the edit was actually made* (**Efficacy** or **Edit success**). De Cao et al. (2021) extended this evaluation to seq2seq models using cloze (fill-in-the-blank) evaluations and introduced two measures: the effectiveness of model edits on paraphrases of input queries (**Generalization**) and how well the model maintains performance on predictions that shouldn't change (**Locality** or **Specificity**) (See Hoelscher-Obermaier et al. (2023) for further explorations of locality). Hase et al. (2021) additionally introduced a measure for understanding the degree to which model editing impacts entailed facts (**Portability**) which was further extended in Cohen et al. (2023). These four measures use the next few tokens after a short prompt to evaluate model editing and are the *status quo* for assessing model editing interventions (Yao et al., 2023). In the paper, these evaluations are called 'short form' as opposed to our 'long form' setting which evaluates paragraph-length texts.

Most contemporary methods of model editing do not consider long-form generation (Hernandez et al., 2023; Huang et al., 2023; Mitchell et al., 2022a,b; Zheng et al., 2023). Meng et al. (2023) introduced an automatic consistency and fluency measure for longer generations based on reference texts from wikipedia and n-gram entropy. However, these were neither validated using human judgements nor fine-grained enough to capture efficacy, generalization, locality, or portability of different model editing techniques on longer form generation. Meng et al. (2023) did perform a preference ranking survey with human raters using fluency, edit success, and factual consistency as ratings. We find that these previous measures do not generally correlate well with human ratings (Appendix I). We build on these preliminary evaluations to establish a more comprehensive view of the impact of model editing on 'long-form' natural language generation.

3 Methods

To measure the quality of model editing in long-form generation, we developed the following measures. These were designed to align with the short-form evaluations. (1) ***Edit consistency*** (is there evidence that the edit was made in a generated passage?) which is intended to align with ***efficacy*** (2) ***Factual consistency*** (are generated passages still consistent with facts that were true before the edit?) which is intended to align with ***locality*** (3) ***Internal consistency*** (the degree to which passages do not contradict themselves or each other) which is aligned with ***portability***, (4) ***Topicality*** (the degree to which the passages stay on topic), and (5) ***Naturalness*** (the fluency of the generated passage). (4) and (5) are intended to measure the impact of model editing on natural language generation (NLG).

We operationalize these by constructing a dataset of prompts for generating highly related passages (§ 3.1), devise a likert scale (§ 3.2) and annotation (§ 3.3) setting that we collect human ratings on and develop automatic measures for (§ 4).

3.1 Coupled Entity Prompts Dataset

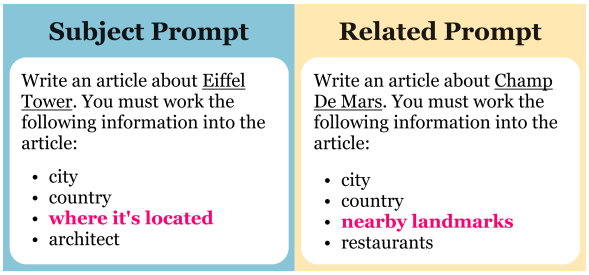

Figure 2: Example prompts used to generate passages. The highlighted property means the other entity is the object of that property

Our dataset, ***Coupled Entity Prompts***, is based on Counterfact (Meng et al., 2023) and zSRE (Levy

et al., 2017). Each sample consists of two highly coupled prompts (see Fig. 2). *Coupling* is the degree to which a related entity shares properties with a subject entity and the ground-truth target from Counterfact or zSRE. For example, Champ De Mars is the park the Eiffel Tower is located at. These entities are highly coupled since they share many properties such as city, country, and near by restaurants. Champ De Mars and Eiffel Tower both share the city of Paris as the ground-truth target which will be updated to Rome for model editing.

The *subject prompt* asks the model to write an article about the subject of an edit (e.g., Eiffel Tower in Fig. 2) and to include a number of properties about that subject (e.g., where it's located). The *related prompt* asks the model to write an article about a related entity (e.g., Champ De Mars in Fig. 2) and its properties where the related entity is highly coupled with the subject.

We define a successful edit in the "long-form" setting as: (1: *Edit consistency*) completing the subject prompt as if the edit is true and the related passage does not contradict the edit, (2: *Factual Consistency*) the subject and related passage minimize changes in the ground truth properties and (3: *Internal Consistency*) the passages should neither contradict themselves nor each other.

We performed a SPARQL query on Wikidata to get related entities that had a relationship to both the subject and pre-edit target (e.g. Paris) for all subject entities in Counterfact and zSRE. We also queried for the ground truth properties about the subject and related entities (e.g. country, city, and restaurants near by). This data was used to construct prompts for a language model to write a paragraph about the subject and related entity and instructed the model to include those ground truth properties so we can measure portability and locality. In total, we constructed 3867 subject and related entity prompts for Counterfact and 3522 for zSRE (see Appendix A for details and examples).

3.2 Evaluation

Likert Scale To measure the questions in § 3, we devised a survey using a 7-point likert scale consisting of nine questions (subject and related passages were rated seperately; internal consistency includes a cross passage consistency rating). See Appendix D for full survey details.

Human ratings To collect human ratings, we randomly chose 12 samples from the Counterfact subset of our dataset. We use three methods to generate two outputs (subject and related passage) from the prompts in § 3.1 for each of these 12 samples. First, we developed a ***No edit*** control setting, where we used the language model `llama2-7b-chat` (Touvron et al., 2023) without making any edit intervention. These samples should rate low on edit consistency and act as a baseline for the other measures. Second, we used the editing method ***ROME*** (Meng et al., 2023) to edit the model and then generate outputs. Finally, the authors of the paper wrote paragraph-length responses to the same prompts to produce a ***human***-written baseline as if the edit were true (see Appendix C for details). We expected the human-written baseline to score highest across all categories. The final result was 72 generated passages (Appendix B.1 shows model generation details).

Sampling The survey was distributed online to eight computer science graduate students who volunteered for the task from a research methods course. The study design included two groups of four participants. In each group, the participants rated the same randomly selected samples. The samples were in a random order to avoid fatigue bias. Each participant rated nine samples (three from each intervention) containing two passages each. In total, we collected 648 ratings for 144 passages (each passage was rated by two raters). See Appendix D for full details.

Automatic Survey Ratings The number of survey ratings was too small for a training set. So we generated synthetic ratings by using 100 held-out samples from our dataset with the following model editing interventions: No edit, ROME (Meng et al., 2023), IKE (Zheng et al., 2023), and FT (Zhu et al., 2020). We generated samples for `GPT-J` (Wang and Komatsuzaki, 2021) and `llama2-7b-chat`. We performed survey ratings using the same survey instructions human participants saw using `GPT-4` resulting in a total of 7,164 ratings (796 per question). Treating this as a training set, we trained `DeBERTav3` large (He et al., 2022) for each question and evaluated the model using the human survey ratings as the test set. For experiments in § 5.2, we train the models on the human survey ratings as well. See Appendix B.3 for details for an overview of how the model was trained and Table 9 for per-

formance details.[1]

3.3 Annotation

For annotations, we present the annotator with a premise which consists of the subject or related entity passage and a claim which consists of one of the following: (1) an edit statement such as "The Eiffel Tower is located in Rome" (2) the pre-edit statement such as "The Eiffel Tower is located in Paris, or (3) a ground truth statement such as "The Eiffel Tower was completed in 1889". Similar to natural language inference, each premise is classified as neutral to, supporting, or contradicting the claim. Additionally, annotators are instructed to highlight sentences that would provide evidence for the classification.

Human Annotations Four authors of the paper performed annotations of 726 premise and claim pairs. We performed a pre-test before annotation, all four authors annotated a sample of 186 premise hypothesis pairs to understand the reliability of our annotation scheme. As measured by Krippendorff's α the annotations had good agreement (α = 0.63). After the pre-test, the four authors split the remaining 540 annotations into two groups of 270 annotations and two annotators annotated each group (α = 0.65). In total, after adjudication for conflicts by a senior author ([REDACTED]), there were 1,496 total classifications and 1,985 evidence sentences collected. See Appendix E for the annotation guidelines.

Automatic Annotations Since we have a large set of human annotations, we finetuned `DeBERTAv3` large (He et al., 2022) on these. We enhanced the dataset with the highlighted sentences and treated those as premises for each claim resulting in a total of 1,642 samples after deduplication. To evaluate this method, we split the human annotations into a train test split of 80% and 20%[2]. See Appendix B.4 for full training details and training set distribution.

4 Experiments

In order to answer our research question of how different model editing interventions compare, we develop a comprehensive suite of experiments that use the automatic measures developed above to evaluate the following interventions: FT with constraint loss (Zhu et al., 2020), MEND (Mitchell et al., 2022a), ROME (Meng et al., 2023), MEMIT (Meng et al., 2022), and IKE (Zheng et al., 2023). We implemented these model editing interventions on `GPT2-XL` (Radford et al., 2019), `GPT-J` (Wang and Komatsuzaki, 2021), `llama2-7b` and `llama2-7b-chat` (Touvron et al., 2023) (the main paper results report `GPT-J` and `llama2-7b-chat` with the other models results in Appendix F). For each model, we also computed a 'no edit' control. We also experimented with using a zero shot `GPT-4` IKE setting (see Appendix B.2) to simulate an upper bound of performance. Subject and related prompts are completed as independent generations.

We perform these evaluations on 100 randomly sampled edits from Counterfact and zSRE. For the zSRE setting, we create two edits per sample. We compute a counterfactual edit, making an edit by changing a true fact to a counterfactual one, as well as a factual edit, changing a false fact to a true one. In total, we assess 300 samples (600 passages).

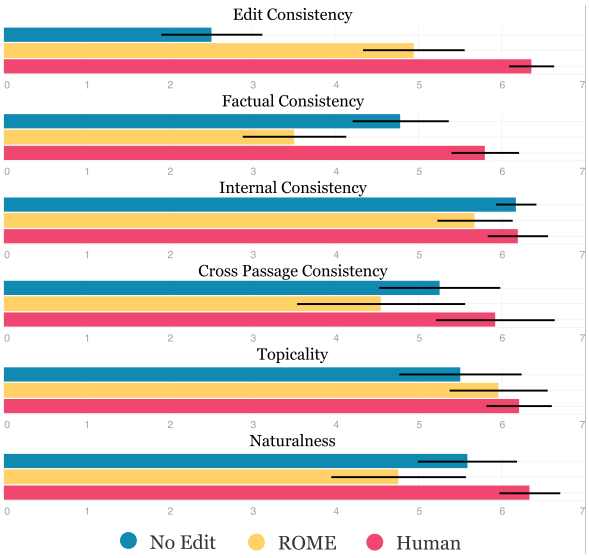

Figure 3: Survey results illustrating the mean rating of long-form quality measures. Human passages always rate highest. ROME is rated even worse than no edit on many dimensions.

5 Results

5.1 Human evaluation

As we would expect (Fig. 3), human written passages were rated higher than all other methods. ROME only approaches human ratings for edit consistency, internal consistency, and topicality. Interestingly the no edit control is rated higher than

[1]We performed additional experiments with zero and few-shot settings using `GPT-3.5`, `GPT-4`, and `llama-2-7b-chat` (see Appendix B.3).

[2]Only a single training run was performed without hyperparameter tuning so a validation split was not needed.

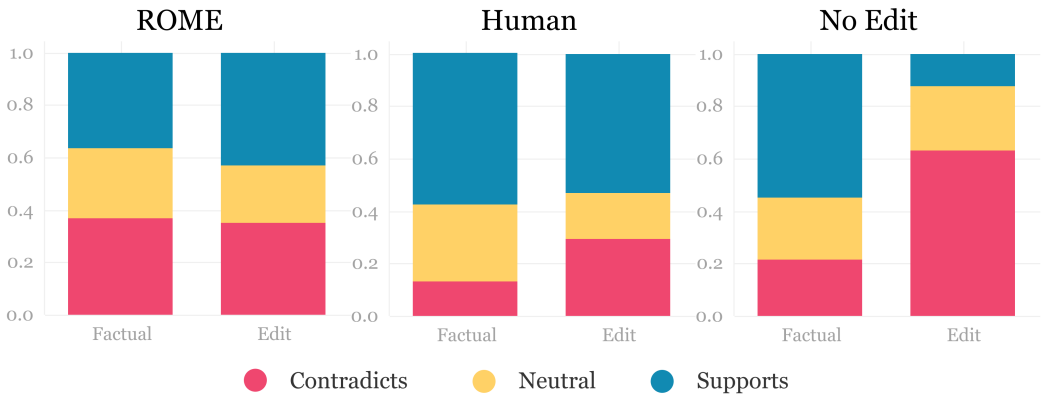

Figure 4: Proportion of labels from human annotation of ROME, human written, and no edit passages. The ground truth is mostly supported in the no edit and human control, while no edit mostly contradicts the edit statements. Human written passages generally are more consistent with the edit statement than ROME passages.

Model	Method	**Edit consistency**		**Factual consistency**		**Internal consistency**			**Topicality**	**Naturalness**
		Subject	Related	Subject	Related	Subject	Related	Cross		
GPT-J	No Edit	1.3±1.3	3.5±1.3	2.3±1.8	3.8±2.3	**6.6±1.5**	**7.0±0.1**	**6.6±1.0**	5.4±2.3	**5.4±2.6**
	IKE	2.0±2.2	3.9±1.6	**2.4±1.7**	**4.1±2.3**	6.4±1.7	6.9±0.4	6.5±1.0	5.4±2.1	5.3±2.6
	FT	1.5±1.7	3.8±1.1	2.1±1.6	**4.1±2.3**	6.5±1.5	6.9±0.7	**6.6±1.0**	**5.6±2.1**	**5.4±2.6**
	MEND	**3.2±2.9**	**4.0±1.8**	**2.4±1.8**	**4.1±2.4**	6.0±2.2	6.7±1.3	6.5±1.2	4.5±2.5	5.1±2.7
	ROME	2.8±2.8	3.9±1.4	1.5±1.1	3.2±2.2	5.9±2.1	6.9±0.6	6.0±1.4	4.1±2.5	4.4±2.9
	MEMIT	2.1±2.3	3.8±1.3	2.0±1.6	3.7±2.2	6.5±1.6	6.9±0.5	6.5±1.1	5.5±2.1	5.3±2.7
llama2	No Edit	2.2±2.4	2.0±1.5	**3.5±1.9**	4.7±2.4	**6.9±0.3**	**7.0±0.1**	**6.6±1.4**	**7.0±0.4**	**6.9±0.8**
	IKE	4.7±2.9	3.4±2.4	3.4±1.9	**4.9±2.2**	6.8±0.6	**7.0±0.2**	**6.6±1.4**	**7.0±0.2**	6.8±1.0
	FT	5.1±2.8	**3.7±2.4**	2.1±1.5	3.7±2.4	5.7±2.4	6.7±1.4	6.1±1.6	5.1±2.7	5.8±2.4
	MEND	3.1±2.9	2.6±1.9	3.4±1.9	4.6±2.3	6.8±0.6	**7.0±0.1**	**6.6±1.3**	6.9±0.5	6.8±1.2
	ROME	**5.4±2.6**	3.5±2.4	1.9±1.4	3.9±2.4	6.4±1.7	**7.0±0.5**	5.8±2.1	6.5±1.5	6.2±2.0
	MEMIT	**5.4±2.7**	3.3±2.3	2.0±1.5	3.8±2.4	6.3±1.8	6.9±0.6	5.9±2.1	6.3±1.8	6.2±2.0
GPT-4	IKE	5.2±2.7	5.1±2.5	3.2±1.9	6.1±1.5	6.7±1.3	7.0±0.0	6.7±1.2	6.7±1.4	7.0±0.0

Table 1: Automatic ratings of zSRE and Counterfact (DeBERTaV3) across editing methods. Significant reduction (one-sided Wilcoxon sign rank, $p < 0.05$) in factual consistency for ROME and MEMIT. llama2 here is llama2-7b-chat .

ROME in almost all dimensions except edit consistency and topicality. This indicates that ROME worsens the general quality of natural language generation. Cross passage consistency is reported separately from other internal consistency measures for illustrative purposes. Ratings were statistically significant (one-sided Wilcoxon sign rank test, $p < .05$) except for no edit and human on internal and cross passage consistency and human and ROME on topicality.

For the annotations, Fig. 4 corroborates our survey findings: both the no edit and human control groups have better factual consistency than ROME as measured by the number of ground truth statements that are supported. Human written passages have better factual consistency and edit consistency than ROME or no edit. All comparisons in between methods were statistically significant (Chi-square test of independence). See Appendix B.4 for annotation distribution details.

5.2 Understanding the impact of model editing across interventions

Table 1 illustrates the quality of various model editing methods using our automatic survey rating approach. Our main findings is that ROME and MEMIT suffer from significant drops in performance on factual consistency[3] and internal consistency (especially cross passage) despite often being the most effective editing method according to short-form evaluations (Appendix H). Except for GPT-4 IKE, there seems to be a pattern where models that do better at edit consistency for the subject passages perform worse on reflecting the edit in the related passages. Unsurprisingly in-context editing (IKE) tends to maintain similar performance to the

[3]We found no statistically significant correlation between factual consistency and edit consistency.

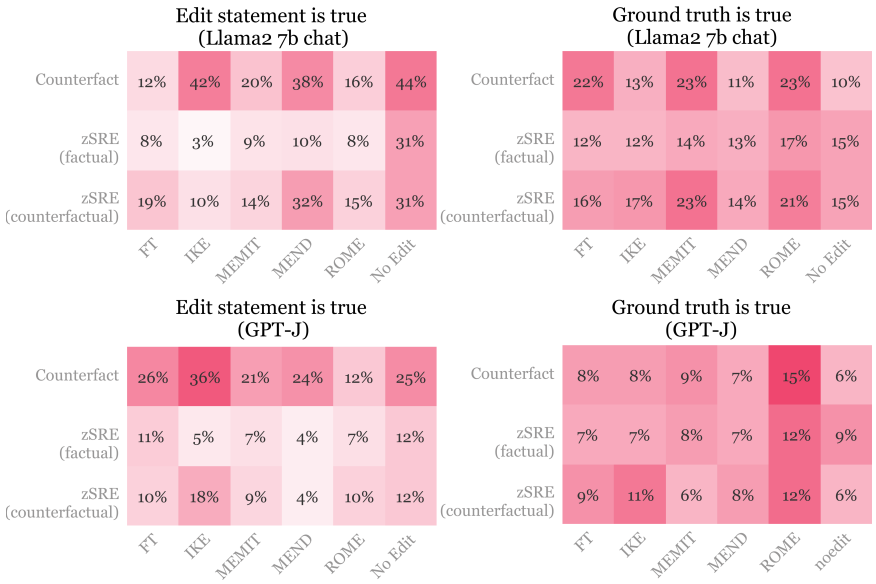

Figure 5: Percentage of claims that contradict the generated passage. Results corroborate our findings that MEMIT and ROME suffer from high factual drift.

'no edit' control across factual consistency, internal consistency, topicality, and naturalness despite not being as effective at edit consistency. Along these lines, `GPT-4` IKE is generally the most effective method especially in the case of maintaining edit consistency and factual consistency in the related passages[4].

5.3 What is the scope of the edit?

Our classifier allows us to *understand the scope of the change introduced by an editing method* since we are able to measure the number of ground truth properties that are contradicted by the generated passage after the edit (see Fig. 5). Importantly, no edit indicates the base level of ground truth or edit statements that would be contradicted before the edit was made. All methods perform better than the no edit control on ensuring the edit statement is not contradicted with particular effectiveness of MEMIT, ROME, and FT on `lama2-7b-chat`. However, MEMIT and ROME introduce a high degree of "factual drift" (suffer from locality problems) since a higher % of ground truth statements are contradicted compared to the no edit control and the other methods. we found no inverse relationship between edit and factual consistency.

5.4 Correlating long- and short-form evaluations?

We only found very weak relationships between the short-form evaluations of edit success, generalization, locality, and portability settings from Yao et al. (2023) and long-form evaluations (Appendix H). Edit consistency generally does capture some of what is measured by the short-form metrics ($\rho \in [0.1, 0.17]$, $p < 0.05$). Cross passage consistency also has weak correlations with portability ($\rho = 0.13$, $p < 0.05$) and generalization ($\rho = 0.12$, $p < 0.05$). Importantly, factual consistency and internal consistency have almost no relationship with short-form measures ($\rho \in [-0.08, 0.05]$, $p < 0.05$). We speculate the reason for this is that the short-form metrics measure superficial token distribution questions about word co-occurrence (see Hoelscher-Obermaier et al. (2023) for an illustration) while our measures require success across much larger generations. Either way, this finding indicates that our evaluation setting measures unique dimensions not captured by short-form evaluation.

5.5 Injection vs updating facts

One limitation with model editing evaluations is that we are not sure if we are updating a previously known fact or injecting a brand new fact since knowing a fact beforehand is model specific[5]. We analyze the performance difference between

[4] Additional comparisons with other models and automatic measures such as zero-shot and simpler baselines are presented in Appendix F.

[5] See a similar analysis in 5.1 of (Hase et al., 2023)

these in Fig. 6 by looking at the mean rating difference on edit consistency and factual consistency measures considering whether *an edit statement was already known* or not (**Edit was already true**), whether the edit is a *counterfactual update* or a novel fact injection (**Counterfactual update**) and whether the edit is *factual correction* of a known but wrong fact or is a novel fact injection. Appendix G illustrates the proportion of samples that represent these categories for each dataset.

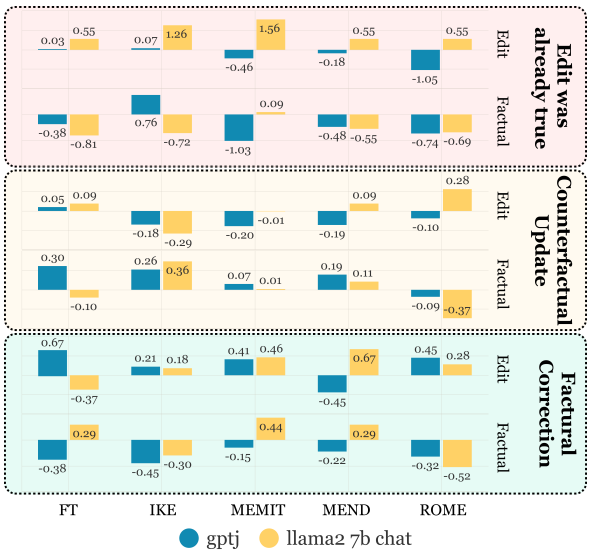

Figure 6: Model performance can differ depending on the type of edit task.

We see a small performance drop on edit consistency if we are doing a counterfactual update rather than a novel fact injection indicating updates are harder than injection (updates only represent 8% and 18% of Counterfact on `GPT-J` and `llama2-7b-chat` see Appendix G). Factual consistency is better for counterfactual updates compared to novel fact injection which might mean that during a novel fact injection, we are simply missing additional necessary ground truth knowledge. For factual correction, we see that generally we do better on edit consistency if we are correcting an erroneous fact. For factual consistency, we do worse in the factual correction setting with some exceptions which means that novel fact injection is easier to maintain ground truth statements on when a model already is biased towards and incorrect answer.

Finally, Appendix G shows how the edit statement is already true in many cases in zSRE. In Fig. 6 we see the implications of this where for `llama2-7b-chat`, if the edit was already true then edit consistency is rated much higher and, as we'd expect since this is a statement that would contradict ground truth, factual consistency is much lower. Overall, these differences aren't large enough to change our results in § 5.2 but we should perform these types of controlled experiments when doing model editing experiments to ensure our results hold across different types of editing tasks[6].

5.6 Error Analysis

In order to understand particular errors made during generation, we manually analyzed 200 samples from Counterfact by selecting the 20 lowest automatically rated samples for each edit intervention for `GPT-J` and `llama2-7b-chat`. Please reference Appendix J for each example we discuss.

First, we found a number of cases of disfluency. Aside from common cases of disfluency in NLG like repetition or completely degenerate generations (often from FT), we found there were cases of nonsensical generations like in Example 1 where Boston gets overused as a noun for categories like profession. Example 2 illustrates a relatively common degenerative case with ROME and MEMIT where space tokens were omitted.

Another common problem was cases with entity or topic drift and lexical cohesion issues. In Example 3 MEMIT correctly edits Paul Guimard's birth place to be in Russia but the change creates a whole new entity with the same name who is a Russian cosplayer born in the 1980s, the Paul Guimard *we intended* to edit stays unedited as reflected in the related passage. Examples 4 and 8 illustrate a common case where the subject entity is introduced at the beginning but the generated passage slowly drifts towards another entity (in this case the Empire Building or IBM Lotus) and continues to drift into another topic. Examples 5 and 7 illustrate cases of poor lexical cohesion where the name of the entity slowly changes over the course of the generation (e.g. Delon becomes Deloy which becomes Deloyg). Another illustrative example from a ROME edit in the human survey is [Benedetto Marcello (1847-1937) was an Italian jazz musician... He was born in Genoa, Italy, to parents Antonino and Teresa Jazz. His family name is Benedetto Jazz] Example 6 is a combination of topic and entity drift where Milan is correctly edited to be located in Japan but the generation drifts towards talking about Milan as if it were an alias for Tokyo and continues referring to

[6]For the readers benefit we present a similar performance analysis for the short evaluations in Appendix H.2

the subject as Tokyo rather than Milan.

Another common case of editing failure is the introduction of contradictions that either contradict with statements made during the main passage (within a single generation) or that conflict with other generations in the related passage. Example 10 states the Ipod was created by Nintendo and then in the next sentence mentions it was created by Apple. Example 13 mentions that Guimard was Groult's cousin but in the related passage they are said to be married. Other edits contradict common sense or world knowledge such as Example 16 where the Dawa River is a river located in Malta but later mentions how the Dawa is a tributary of the Jubba River which the model says is in Somalia.

Finally, reflecting our finding that some models tend to violate more ground truth properties than others, we found success cases where some models only made minimal edits (Example 17) or edits that incorporate both the edit statement and the pre-edited fact (Example 11), while other edits introduced very large changes violating locality such as Example 14 where changing Jeanne Moreau's birth place to Poland unnecessarily changes her teacher Denis d'Inès to be Polish as well when generating the related passage (a reflection of poor locality). Again Example 15 does not just change the band Barren Earth's location to Sydney, Australia but also changes the subgenre of the band as well as the members of the band. While IKE is generally an effective method for editing larger models can reject the edit. Example 20 illustrates a case with `GPT-4` IKE where the edit is rejected by the model.

6 Discussion

Current model editing methods have many gaps that are not measured by short-form evaluation methods and the preliminary 'long-form' methods from (Meng et al., 2023) don't correlate well with human data. Factual drift, where methods like ROME tend to make much larger changes than MEND or FT is not revealed in the standardized 'shot-form' measures from Yao et al. (2023). Future efforts should be devoted to balancing factual drift and edit success in NLG.

Factual drift might be a desirable feature of model editing, where there are model edits that should imply changes that would contradict ground truth statements. However, we want to develop evaluation methods that are able to measure the trade off between edit and factual consistency which we believe our methods are able to measure. RippleEdit (Cohen et al., 2023) is a good step in this direction for short-form evaluation that could inspire future work on more comprehensive long form evaluations that measure the scope of change beyond a single related passage.

Finally, our results reveal an important general property we should be looking for in high-quality model editing methods: consistency. The problematic generations that we investigated often indicated cases of contradiction, whether that was self contradiction, contradicting separated generations in the related passages, or contradicting ground truth statements. As developers of model editing interventions, we should design methods that result in generations that have high consistency: there should at least be no contradictions across generated passages. For cases where we allow a high factual drift, we still want to ensure self consistency. Other properties like fluency and topicality are important properties which tend to suffer and we should ensure that novel methods do not inadvertently harm general NLG quality.

7 Conclusion

In this paper, we introduced two automatic methods, survey ratings and classification, for evaluating the impact of model editing on natural language generation in paragraph-length generation settings. We validated these measures by collecting survey and annotation data from human participants and then developed a trained model setting that correlated well with human data.

Using these automatic metrics, we performed a comprehensive analysis of the natural language generation quality of common model editing techniques finding the following results: (1) ROME and MEMIT suffer from a high factual drift from ground truth statements compared to other methods like MEND or IKE (2) there is very little relationship between previous short form evaluations like generalization, locality, and portability with our long form metrics (3) through a qualitative study, we presented a number of common failure modes such as entity drift, lexical cohesion, internal contradiction, and scope errors. We hope that identifying these failure modes can help the community develop future model editing techniques that work well in "long-form" settings.

8 Limitations

The primary limitation of our study is the small sample size and weak inter-rater reliability of our survey filled out by human participants. It is important to note that we initially had a Krippendorff α of .3 and reran the survey with a new of set of participants but only marginally increased the agreement to .34[7]. The surveys took on average one hour to complete and is much more laborious to complete than the annotations. To further develop the survey method we should investigate ways of increasing both the inter-rater reliability and efficiency.

We performed the study using a limited demographic of graduate computer science students who would be familiar with the language of natural language generation. Studies looking to scale up our method with diverse demographics such as from crowdsourcing would likely suffer from even worse agreement. One alternative could be finding alternative ways to operationalize measures like internal consistency and topicality.

Another limitation is that our methods only implicitly captures generalization, locality, and portability so we can't speak directly to specific effects on these properties with our measure. Related, the study only uses one related entity when generating and assessing our related passage. To further assess the scope of impact, future methods should incorporate generated passages farther away in the knowledge graph than the highest coupled entities.

One notable gap in our study that should be followed up on is the question of the impact of batch and chained editing has on NLG quality. Since we can imagine many settings in which a user would want to make a large amount of edits to a language model or make subsequent edits one after another, we would want to understand what impact that has on NLG separately from short evaluations.

Finally, it is important to acknowledge the ethical concerns with using counterfactual editing datasets. These datasets purposely introduce misinformation to determine the efficacy of an editing technique. As a community we should be aware that a side effect of this research could be demonstrating comprehensive methods for injecting misinformation and as such we should look towards moving away from counterfactual editing towards factual correction datasets or datasets that have less misinformation harm risk such as edits in fictional settings.

[7] Agreement on some measures like edit consistency were much higher ($\alpha = 0.55$) see Appendix D.1.

References

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the Ripple Effects of Knowledge Editing in Language Models. ArXiv:2307.12976 [cs].

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing Factual Knowledge in Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506. Conference Name: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing Place: Online and Punta Cana, Dominican Republic Publisher: Association for Computational Linguistics.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs. ArXiv:2111.13654 [cs].

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and Editing Knowledge Representations in Language Models.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11548–11559, Toronto, Canada. Association for Computational Linguistics.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-Patcher: One Mistake worth One Neuron. Publisher: arXiv Version Number: 1.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, pages 1–33.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and Editing Factual Associations in GPT. ArXiv:2202.05262 [cs].

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-Editing Memory in a Transformer.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast Model Editing at Scale. ArXiv:2110.11309 [cs].

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-Based Model Editing at Scale. ArXiv:2206.06520 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. ArXiv:2305.13172 [cs].

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can We Edit Factual Knowledge by In-Context Learning? ArXiv:2305.12740 [cs].

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying Memories in Transformer Models. ArXiv:2012.00363 [cs].

A Dataset Construction Details

The following SPARQL query was used to select the related entities for each subject entity and ground truth target. The query counts and orders by the number of shared properties to select the related entity with the highest degree of coupling.

```
SELECT ?item
WHERE {{
  {{ ?item ?p wd:{subject} . }}
  UNION
  {{ wd:{subject} ?p ?item . }}
  {{ ?item ?p2 wd:{target} . }}
  UNION
  {{ wd:{target} ?p2 ?item . }}
}} GROUP BY ?item
   ORDER BY DESC(COUNT(?item))
   LIMIT 1
```

We collected 3867 samples using the Counterfact dataset as our source of subject entities and ground truth targets and 3522 samples from zSRE. Appendix A.1 illustrates what the dataset looks like. We construct the prompts using the property keys of the ground truth without the values.

A.1 Dataset Example

Below is an example from our dataset that illustrates the ground truth statements and the resulting subject and related prompts.

```
{
    "coupled_prompts_and_properties": {
        "subject_entity": {
            "ground_truth": {
                "country of citizenship
                    ↪ ": [
                    "Philippines"
                ],
                "occupation": [
```

```
                    "politician",
                    "engineer"
                ],
                "child": [
                    "Ramon Magsaysay, Jr
                        ↪ ."
                ]
            },
            "entity": "Ramon Magsaysay"
        },
        "coupled_entities": [
            {
                "entity": "Ramon
                    ↪ Magsaysay, Jr.",
                "subject_as_object": [
                    "father"
                ],
                "ground_truth": {
                    "occupation": [
                        "politician"
                    ],
                    "country of
                        ↪ citizenship":
                        ↪  [
                        "Philippines"
                    ],
                    "father": [
                        "Ramon Magsaysay
                            ↪ "
                    ]
                }
            }
        ]
    }
}
```

Subject Prompt:

```
Write an essay about Ramon Magsaysay
Include the following information:
- manner of death
- place of burial
- child
- member of political party
- country of citizenship
- position held
- occupation
```

Related Prompt:

```
Write an essay about Ramon Magsaysay, Jr
Include the following information:
- father
- country of citizenship
- position held
- sex or gender
- occupation
```

B Models used and generation details

B.1 Model generation for samples details

For all of the generation settings used for generating outputs after each model editing intervention we used the following parameters: top k of 50, top p of 0.95 and temperature of 0.9. For `GPT-3.5` and `GPT-4`, we used a temperature of 1. We initially attempted greedy sampling for generating outputs after model editing but the scores were much worse. `gpt-3.5-turbo-0613` is used for `GPT-3.5` and `gpt-4-1106-preview` is used for `GPT-4`.

B.2 GPT-4 IKE

In order to operationalize IKE for `GPT-4` we used the following prompts to generate passages based on the human-written instructions in Appendix C. Unfortunately using those instructions directly led to the model to often refuse to incorporate the edit. The prompts below are what we eventually settled on after seeing a reduction in refusal behaviour. We used this model zero-shot without additional exemplars of edits as it was determined as not needed to complete the task and few shot IKE would introduce more cost to an already expensive generation setting. Finally, it's important point out that we do not know the extend to which zSRE or Counterfact datasets are incorporated into the `GPT-4` training set.

Instruction Prompt:

```
As an AI system, some of your factual
    ↪ knowledge is incorrect or
    ↪ outdated.
we will provide edits that correct old
    ↪ outdated facts that are incorrect
    ↪  in your system.
You will be given a prompt with an edit
    ↪ and a passage and asked to write
    ↪ a new passage taking the factual
    ↪ correction edit into account.
```

Write essay Prompt:

```
Given an edit write an updated passage
    ↪ to reflect this new knowledge.
    ↪ Include discussion on all the
    ↪ properties given as well as the
    ↪ edit.
Edit: {edit}
Prompt: {prompt}
```

B.3 Automatic survey ratings

We develop a zero and few-shot setting where we prompt a large language model with the same guidelines and instructions that humans received during the survey. For the few-shot settings we randomly sample demonstrations from the human surveys for the question being answered excluding demonstrations from the sample that is currently being evaluated. We report results from this method on `GPT-3.5`, `GPT-4`, and `llama-2-7b-chat`.

The prompts used in the zero- and few-shot set-

tings are the same as the questions in Appendix D. Generally, the full guidelines and instructions do not fit into the token space and do not allow a few-shot settings with several demonstrations. In order to fit the prompt in the token space, we only present relevant instructions to one survey question at a time.

We trained three sets of the nine rater models fine-tuned on the dataset described in § 3.2. The first setting contains none of the human ratings in the training dataset achieving Krippendorff's α of 0.45. In the second, we use half of the human ratings and keep the other half as the held out set for evaluation. This is the model used in the agreement measures in Table 9. Finally, for the automatic ratings presented in § 5.2 we train the model on all human ratings which has Krippendorff's α of 0.62. For training we finetune `DeBERTaV3` large using the following hyperparameters:

```
learning_rate=6e-6
batch_size=1
train_epochs=20
weight_decay=0.01
warmup_steps=1000
gradient_accumulation_steps=4
fp16=True
```

The training of these models took place using *Digital Research Allaince of Canada's* infrastructure. We used 4 A100 GPUs with 40GB vRAM and 4 V100 GPUs with 32GB vRAM.

B.4 Classification

For annotation, similar to the automatic survey evaluation we develop a zero- and few-shot setting. The prompts also use the same guidelines and instructions that annotators received. The few-shot setting samples from the human annotations excluding the sample being presented to the model for evaluation. For the `DeBERTaV3` large model, we do not collect highlighted sentences.

The prompts used in the zero and few-shot settings present the guidelines in Appendix E and a claim premise pair.

The classification model is trained on the dataset described in § 3.3 using `DeBERTaV3` large with the same hyperparameters and compute as above. The performance is reported in Appendix F. The distribution of annotations used during training in presented in Appendix B.4 and Appendix B.4.

Classification	Proportion
Contradicts	22%
Neutral	37%
Supports	41%

Table 2: Distribution of annotations according to annotation label.

Classification Setting	Proportion
Ground truth and main passage	61%
Ground truth and related passage	19%
New fact and main passage	5%
New fact and related passage	5%
Old fact and main passage	5%
Old fact and related passage	5%

Table 3: Distribution of annotations according to type. New fact refers to the edit statement; old fact refers to the pre-edit statement.

C Human Written Edit Details

The following are the instructions given when producing the human-written edit texts. When presented with the prompt, the participants were also given the ground truth about what was true about the properties for the subject and related entity.

D Survey Instrument

The survey was constructed using a google form for each participant. The survey was distributed via email to participants who agreed to participate. Informed consent was given in the instructions of the survey and in the volunteer solicitation process, participants were allowed to opt out at any time. Participants were not compensated for filling out the survey. The survey took an average of 1 hour to complete. Instructions can be found in Table 6.

The inter-rater reliability (Krippendorff) between each group in the survey was weak (α = 0.34) indicating a high degree of subjectivity in the task at hand (some question types like Edit consistency are higher (α = 0.55) see Appendix D.1).

D.1 Inter-rater reliability broken out

Appendix D.1 shows how high agreement was dominated by edit consistency, factual consistency and Naturalness. Our survey measures of cross passage and internal consistency generally had poor agreement meaning they were generally not understood by the survey participants. This is reflected in our main results for internal consistency for the auto-

Instructions:
Given an edit write two paragraphs:
(1) An informative article about the subject as if the edit was true. Include discussion on all the properties given.
(2) An informative article about the related entity as if the edit was true. Include discussion on all the properties given.
Example:
Edit: The genre played by Olu Dara is opera
Subject prompt: Write a short paragraph about Olu Dara
You must work the following information into the article:
- place of birth
- country of citizenship
- occupation
- religion or worldview
- instrument
- genre
- given name
- record label
- child
- sex or gender
- relationship to In the World: From Natchez to New York

Paragraph: Olu Dara was an opera singer and jazz musician who was born in Natchez, Mississippi. He played a variety of instruments but is most known for his jazz guitar and trumpet skills. In opera, he was most known for his role as the lead in the opera In the World: From Natchez to New York as a tenor. He was also the father of the famous rapper Nas. He is a member of the Nation of Islam and is a devout Muslim.

Related Prompt: Write a short paragraph about In the World: From Natchez to New York
You must work the following information into the article:
- performer
- genre
- record label
- relationship to Olu Dara

Paragraph: In the World: From Natchez to New York is a jazz opera written by Olu Dara. It was released in 1998 by Atlantic Records. It was performed by Olu Dara and his son Nas. It was produced by Olu Dara and Nas for Atlantic Records.

Table 4: Instructions for human participants to write passages as if the edit were true including two example paragraphs of what was written.

Question Type	α
Edit consistency	0.55
Factual consistency	0.21
Naturalness	0.21
Topicality	0.09
Cross passage consistency	0.01
Internal consistency	-0.02

Table 5: Inter-rater reliability between participants taking our survey broken down by question type. Topicality, cross passage consistency, and internal consistency have quite poor inter-rater reliability. We don't feel this invalidates our study due to the high subjectivity of the task but it does speak to improvements that should be made for internal consistency measures in particular.

matic ratings which don't illustrate anything very interesting.

Survey Instructions

AI Text Generation Fact Changing Survey

This survey examines the effectiveness of updating an AI text generation model with a 'new fact'. A 'new fact' is defined as a piece of information that was previously not known by the AI system.

All data collected here will be entirely anonymous. By filling out this survey you are consenting to the public sharing of anonymized raw data of this survey for the purposes of reproducability as well as constructing a dataset to help improve future AI systems. You may opt out of the survey at anytime.

Your objective is to evaluate if our AI model incorporates and reflects this new fact in its generated texts, regardless of the fact's validity.

Note that these 'new facts' might not be widely recognized as truthful. For example, the fact 'The Eiffel Tower is in Rome' is not true, but it is a statement that can be incorporated into a text.

We'll present a 'new fact' along with two AI-generated passages:

- one about the subject of the fact (the main passage).

- another about a related entity (the related passage).

In the example 'The Eiffel Tower is in Rome'

- the subject is 'The Eiffel Tower'

- A related entity is 'Champ de Mars' (a location the Eiffel Tower is near)

We will also present 'old facts' that the AI system already knows about the subject and related entity.

Some of these may change as a result of the new fact. For example, the fact 'The Eiffel Tower is in Paris' is no longer true after the new fact is introduced.

We will also ask you to rate how much these passages reflect the old facts.

Remember, your role is not to assess the truthfulness of the fact. Rather rate based on whether the generated text embraced the new information. You will be asked to rate, on a scale of 1 (Strongly Disagree) - 7 (Strongly Agree) several questions about 9 samples. **Some of the samples you will be asked to evaluate may be very similar or the same - this is on purpose!** Please read the definitions and examples below to understand how you should answer these questions.

Definitions

Main Passage: The passage about the subject of the new fact

Related Passage: The passage about a related entity to the subject

New Fact: A piece of information that was previously not known by the AI system

Old Fact: A piece of information that was previously known about the subject or related passage The old fact may change as a result of the new fact.

Subject: The subject of the new fact

Related Entity: The related entity to subject (for example father, mother, brother, sister, etc.)

Consistent: The degree to which the passage supports or does not contradict the new fact, itself, or the other passage

Examples:

1. There is evidence the new fact is true in the main passage

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Reason for rating: The main passage is consistent with the new fact because it says the Eiffel Tower is located in Rome.

Negative Example (Rating of Strongly Disagree):

Main passage: The Eiffel Tower, located in Paris, France, is one of the most iconic landmarks in the world and has become a symbol of French culture and engineering prowess.

Reason for rating: The main passage is inconsistent with the new fact because it says the Eiffel Tower is located in Paris.

2. There is evidence the new fact is true in the related passage

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Related passage: The Champ de Mars is a large public greenspace in Rome, Italy, located near the Eiffel Tower.

Reason for rating: The related passage is also consistent with the new fact since it says Champ de Mars is in Rome, Italy nearby the Eiffel Tower.

Negative Example (Rating of Strongly Disagree):

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The related passage is inconsistent with the new fact since it says the Eiffel Tower is in Paris and located near the Champ De Mars which is also in Paris.

3. The main passage is consistent with the old facts

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower completed in 1887, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Old fact: The Eiffel Tower was completed in 1887.

Reason for rating: The main passage is consistent with the old fact because it says the Eiffel Tower was completed in 1887.

Negative Example (Rating of Strongly Disagree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Old fact: The Eiffel Tower is located in France.

Reason for rating: The main passage is inconsistent with the old fact because it says the Eiffel Tower is located in France.

4. The related passage is consistent with the old facts

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Old fact: The Champ de Mars is in Paris.

Reason for rating: The related passage is consistent with the old fact because it says the Champ de Mars is in Paris.

Negative Example (Rating of Strongly Disagree):

Related passage: The Champ de Mars is situated in Rome.

Old fact: The Champ de Mars is in Paris.

Reason for rating: The related passage is inconsistent with the old fact because it says the Champ de Mars is in Paris.

5. The main passage is consistent with itself

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Reason for rating: the main passage is consistent itself

Negative Example (Rating of Strongly Disagree):

Main passage: The Eiffel Tower was built in Rome in 1887. It was overseen by Gustave Eiffel, a French engineer and architect who was born in 1832 and passed away in 1903 as well as Giovanni Battista Piranesi who was born in 1720 and died in 1778.

Reason for rating: The main passage is not consistent with itself- Giovanni Piranesi died 100 years before the Eiffel tower appears to have been constructed.

6. The related passage is consistent with itself

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The related passage is consistent with itself since there are no contradictions.

Negative Example (Rating of Strongly Disagree):

Related passage: The Champ de Mars is situated in Rome. The large public greenspace is a popular tourist attraction in Paris.

Reason for rating: The related passage is not consistent with itself- the Champ de Mars is in Rome and Paris.

7. The passages are both consistent with each other

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Related passage: The Champ de Mars is situated in Rome near the Eiffel Tower. Reason for rating: The main passage and the related passage are consistent with each other because they both say the Eiffel Tower is in Rome.

Negative Example (Rating of Strongly Disagree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The main passage and the related passage are not consistent with each other because the main passage says the Eiffel Tower is in Rome and the related passage says the Eiffel Tower is in Paris.

8. The main passage is focused on the subject and the related entity is focused on the related entity

For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The main passage is about the subject and the related passage is about the related entity. Neither of the passages drift away from what they are supposed to be about.

Negative Example (Rating of Disagree):

Main passage: Rome is full of great restaurants and shopping. Rome is an amazing place to visit.

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The main passage isn't about the Eiffel Tower at all but the related passage is about the related entity.

9. Both passages are natural sounding text close to what a human would write. For the new fact: The Eiffel Tower is in Rome

Positive Example (Rating of Strongly Agree):

Main passage: The Eiffel Tower, located in Rome, Italy, is one of the most iconic landmarks in the world and has become a symbol of Italian culture and engineering prowess.

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower and the Seine River.

Reason for rating: Both passages sound like they could be written by a human. **Negative Example (Rating of Disagree):**

Main passage: Eiffel Tower Eiffel Tower Eiffel Tower Eiffel Tower Eiffel Tower Eiffel Tower. The Eiffel Tower is in Rome. r ome is fullofgreat restaurants and shopp amazingplacetovisit.

Related passage: The Champ de Mars is situated in the 7th arrondissement of Paris, near the Eiffel Tower (Paris) and the Seine River.

Reason for rating: The main passage has many repetitions, grammar mistakes, and various typos and other errors but the related passage seems fine.

Table 6: Survey instructions that were given to participants.

Edit Consistency:
The main passage is written as if the new fact is true
The related passage does not contradict the new fact
Factual Consistency:
Ignoring the new fact, most of the old facts are still true in the main passage.
Ignoring the new fact, most of the old facts are still true in the related passage.
Internal Consistency:
Ignoring the old and new facts, the main passage does not contradict itself.
Ignoring the old and new facts, the related passage does not contradict itself.
Ignoring the old and new facts, the main passage and the related passage do not contradict each other.
Topical Cohesion
The main passage is focused on the subject and the related passage is focused on the related entity
Fluency
Both passages are natural sounding text close to what a human would write.

Table 7: The questions we used in our survey. Each question was accompanied with a 7 point graphical rating scale ranging from strongly disagree, disagree, somewhat disagree, neutral, agree, somewhat agree, strongly agree.

D.2 Survey Questions

To answer questions about ***Edit Consistency*** we asked participants to rate: "The main passage is written as if the new fact is true" and "The related passage does not contradict the new fact." To capture ***Internal Consistency*** we asked participants to rate: "Ignoring the old and new facts, the main passage does not contradict itself" and "Ignoring the old and new facts, the related passage does not contradict itself". For ***Cross passage consistency***: "Ignoring the old and new facts, the main passage and the related passage do not contradict each other." For ***Factual Consistency*** we asked: "Ignoring the new fact, most of the old facts are still true in the main passage" and "Ignoring the new fact, most of the old facts are still true in the related passage." In addition to consistency properties we also have a question about ***Topicality***: "The main passage is focused on the subject and the related passage is focused on the related entity" and ***Naturalness***: "Both passages are natural sounding text close to what a human would write." See Appendix D for the instructions provided to participants as well as an example sample for rating. These questions are summarized in Table 7.

E Annotation Guidelines

Table 8 presents the annotation guidelines that were given to annotators to read before annotation. All annotations were done using the light tag platform (Perry, 2021).

Annotation Instructions

In this task you will read a passage of text and a claim about that passage in the form of a sentence. You have two jobs:

(1) Classify the passage as supporting, contradicting, or neutral towards the claim.

(2) Highlight the sentences that support or contradict the claim (if the claim is supported or contradicted).

For (2) highlight entire sentences. Try to highlight as many sentences as possible that support or contradict the claim. You may highlight more than one sentence if it captures the context needed or provides additional support or contradiction.

Example of supporting passages

A supporting passage means there is direct evidence for (or in support of) the claim in the passage. If there is some evidence for the claim but not total evidence you should still consider it supporting.

Example:

Passage: Rome is home to the world famous Eiffel Tower. Rome is a great tourist destination and has incredible food. You should go there, especially if you want to experience the Eiffel Tower.

Claim: The Eiffel Tower is in Rome.

Label: supports

Highlighted sentences: Rome is home to the world famous Eiffel Tower. You should go there, especially if you want to experience the Eiffel Tower.

Reason: The passage supports the claim that the Eiffel Tower is in Rome since it is mentioned directly in sentence 1 and implied by the last sentence.

Example of contradicting passages

A contradicting passage means there is direct evidence against the claim in the passage. If there is partial support but the passage contradicts even a little, please consider it contradicts.

Example:

Passage: Rome is home to the world famous Eiffel Tower. Rome is a great tourist destination and has incredible food. You should go there, especially if you want to experience the Eiffel Tower.

Claim: The Eiffel Tower is in Paris.

Label: contradicts

Highlighted sentences: Rome is home to the world famous Eiffel Tower. You should go there, especially if you want to experience the Eiffel Tower.

Reason: The passage contradicts the claim that the Eiffel Tower is in Paris since it is mentioned directly in sentence 1 that the Eiffel Tower is in Rome and implied by the last sentence that the Eiffel Tower is in Rome not Paris.

Example of a neutral passage

A neutral sentence pair is a pair of sentences that neither contradict or support each other. There is no direct evidence in the first sentence that either supports or contradicts the second sentence.

Example:

Passage: Rome is home to the world famous Eiffel Tower. Rome is a great tourist destination and has incredible food. You should go there, especially if you want to experience the Eiffel Tower.

Claim: The Eiffel Tower was built by Gustave Eiffel

Label: contradicts

Highlighted sentences: None

Reason: There is nothing that either contradicts or supports the claim that the Eiffel Tower was built by Gustave Eiffel

Table 8: The instructions used to guide annotators.

Model	Survey			Annotations		
	α	ρ	abs	w/ 1	α	accuracy
llama2 (8 shot)	0	0	19%	79%	0.21	42%
llama2	0.01	-0.01	21%	76%	-0.03	42%
GPT 3.5 (8 shot)	0.22	0.25	27%	87%	0.44	57%
GPT-3.5	0.33	0.28	31%	77%	0.45	54%
GPT-4	0.34	0.28	33%	81%	0.57	69%
GPT 4 (8 shot)	0.49	0.37	31%	84%	0.61	72%
DeBERTaV3	0.59	0.48	42%	85%	0.8	85%

Table 9: Agreement between large language models performing the survey or annotation task and humans performing the task showing moderate agreement for the largest models on the survey and strong agreement on the annotation task. The trained models perform better than zero or few-shot settings.

F Additional Automatic Measures

Table 9 illustrates the degree to which our proposed automatic measures agree with human data. We report Krippendorff's α, Spearman's ρ, absolute agreement (abs), agreement within 1 (w / 1), and accuracy. At first glance, these measures seem to have weak to moderate agreements but when we consider that inter-rater reliability for the survey was low (α = 0.34) for the survey and moderate for the annotations (α = 0.65), we see that GPT-4 approaches these scores especially under an few-shot setting with 8 exemplars. GPT-3.5 is not far behind but llama2-7B-chat is not able to achieve an acceptable rate of agreement even in a few-shot settings. The most promising automatic measures are based on the DeBERTaV3 large models and so we use these to report the test of our results (Due to cost considerations with GPT-4, only GPT-3.5 results are reported in the Appendix F). Appendix F largely corroborates our findings of 'factual drift' present in ROME and MEMIT versus other methods.

F.1 Evaluations broken out by dataset

For the readers benefit we also present the evaluations using the DeBERTaV3 large rating model broken out by dataset and incorporating GPT2-XL and llama2-7b. First, these results illustrate why llama2-7b-chat was chosen over llama2-7b to present results in the main section: the performance is generally much better. We should note that for counterfactual editing in Counterfact (Appendix F.1) and zSRE (counterfactual) (Appendix F.1), GPT-4 IKE is not as effect of an editing method as ROME and MEMIT on llama2-7b-chat. For factual correctness updating with zSRE (factual) in Appendix F.1, we point out the difference between No Edit and other methods, which illustrates the general efficacy of the factual correction subtask of model editing.

G Performance Analysis

Appendix G presents the proportion of samples that are either counterfactual updates (Edit is a fact update (%) for Counterfact or zSRE (counterfactual)), factual updates (Edit is a fact update (%) for zSRE (factual)) or if the edit statement was already true before making the edit.

H Short Evaluations

We replicated the short evaluations presented in Yao et al. (2023) in Appendix H, Appendix H, and Appendix H. For each intervention, we also measured the short-form evaluation settings of efficacy, generalization, locality, and portability using the same evaluation setting as Yao et al. (2023). In addition, we also added an additional short evaluation scenario: whether or not the pre-edit statement such as "The Eiffel Tower is in Paris" is true before the edit. This allows us to understand if we are changing a previously known fact or teaching the model a brand new fact. For the tables below we report how often "ground truth" remains true after the edit; we find that in many cases the "ground truth" tokens can be true in many cases where the edit was successful.

H.1 Correlation with long-form measures

In Appendix H we present the statistically significant ($p < 0.05$) positive Spearman's rank correlations between long-form and short-form metrics with correlation above 0.1. Interestingly Factual and Internal consistency have statistically significant correlations around 0 indicating they are generally not measured by short-form measures.

H.2 Performance Analysis on Short Evaluation

Similar to our performance analysis on the long-form evaluation, we also performed a performance analysis on the short form evaluations in Fig. 7. Ground truth was true corresponds to counterfactual updates. Given that the scores are out of 1, there can be quite large differences for example on FT where there is up to 40% better success if the edit was already true beforehand or we are

Model	Method	Edit consistency		Factual consistency		Internal consistency			Naturalness	Topicality
		Subject	Related	Subject	Related	Subject	Related	Cross		
GPT2-XL	No edit	1.6±1.4	2.8±1.8	3.6±2.4	**3.5±2.4**	**6.3±1.3**	**6.6±0.8**	**5.1±1.6**	3.8±2.3	4.5±2.5
	IKE	**2.6±2.4**	3.2±2.0	3.5±2.5	2.8±2.2	**6.3±1.4**	6.3±1.5	4.8±1.9	4.1±2.3	**4.9±2.4**
	FT	1.8±1.6	**3.3±1.9**	3.2±2.3	3.0±2.2	6.0±1.5	6.4±1.2	5.0±1.8	3.4±2.2	4.3±2.6
	MEND	1.7±1.4	3.1±1.8	**3.7±2.3**	**3.5±2.5**	**6.3±1.3**	6.0±1.6	4.8±2.0	**4.2±2.3**	4.7±2.5
	ROME	2.5±2.4	3.0±1.7	2.7±2.2	2.4±2.0	5.9±1.8	6.1±1.6	4.4±1.9	2.7±2.2	3.8±2.7
	MEMIT	2.1±1.9	3.0±1.9	3.2±2.4	3.2±2.1	6.1±1.5	6.4±1.1	4.9±1.9	3.7±2.3	4.5±2.6
GPT-J	No edit	1.7±1.5	3.2±1.8	4.5±2.3	3.7±2.4	6.6±0.9	6.5±1.1	5.2±1.8	4.9±2.2	**5.4±2.3**
	IKE	2.2±2.1	**3.4±2.1**	4.0±2.5	3.6±2.4	**6.7±0.6**	6.2±1.4	**5.6±1.7**	4.5±2.4	5.0±2.4
	FT	1.9±1.8	**3.4±2.0**	4.5±2.2	3.5±2.4	6.6±0.8	6.4±1.3	5.2±1.8	4.9±2.3	**5.4±2.2**
	MEND	2.1±2.1	3.2±2.1	**4.8±2.3**	**3.9±2.4**	6.5±0.9	**6.6±1.1**	5.2±1.9	**5.0±2.1**	**5.4±2.2**
	ROME	**2.5±2.4**	3.1±2.1	2.3±1.9	2.6±2.1	5.4±2.0	5.6±2.0	4.4±1.9	2.3±1.8	2.9±2.5
	MEMIT	**2.5±2.2**	3.2±1.9	3.7±2.4	3.7±2.3	6.3±1.1	**6.6±1.0**	5.1±1.8	4.0±2.4	5.0±2.4
llama2-7b	No edit	1.8±1.6	4.1±2.1	**5.5±2.0**	**4.5±2.5**	**6.7±0.7**	**6.7±0.8**	**5.9±1.4**	**6.0±1.4**	**6.3±1.4**
	IKE	2.8±2.4	**4.3±2.2**	**5.5±1.9**	**4.5±2.4**	6.5±1.0	6.6±1.1	5.6±1.6	5.6±1.9	6.1±1.6
	FT	**4.3±2.7**	4.0±2.2	3.4±2.4	2.9±2.3	5.7±1.8	5.7±2.1	5.0±2.0	3.7±2.4	4.5±2.5
	MEND	2.3±2.1	3.6±2.0	5.4±1.9	4.4±2.4	6.6±0.9	6.6±0.7	**5.9±1.4**	5.9±1.5	6.1±1.7
	ROME	3.3±2.7	3.4±2.0	2.8±2.3	3.2±2.3	5.8±1.9	6.5±1.2	5.2±1.7	3.5±2.4	4.4±2.7
	MEMIT	3.4±2.7	4.1±2.2	2.6±2.2	3.3±2.4	5.5±1.9	6.2±1.6	4.7±2.1	2.8±2.1	4.5±2.6
llama2-7b-chat	No edit	2.3±2.1	4.2±1.9	5.7±1.8	5.0±2.2	**6.9±0.3**	**6.9±0.3**	6.4±1.0	**6.7±0.5**	**6.7±0.7**
	IKE	2.6±2.6	4.2±2.3	5.8±1.8	4.8±2.2	**6.9±0.4**	**6.9±0.3**	**6.5±0.7**	6.6±0.5	**6.7±0.6**
	FT	**5.1±2.6**	**4.5±2.3**	3.2±2.6	3.6±2.5	6.0±2.0	6.5±1.5	5.7±1.9	5.3±2.0	6.0±1.8
	MEND	2.4±2.4	3.9±2.2	**5.9±1.6**	**5.4±2.2**	**6.9±0.4**	**6.9±0.4**	6.4±0.9	**6.7±0.5**	**6.7±0.5**
	ROME	5.0±2.7	4.3±2.2	3.3±2.5	3.9±2.4	6.6±1.3	6.8±0.6	5.8±1.8	5.9±1.6	6.2±1.5
	MEMIT	4.0±2.8	4.2±2.0	2.8±2.4	3.9±2.5	6.4±1.5	6.7±1.0	5.8±1.8	5.8±1.9	6.3±1.6

Table 10: Survey Ratings by GPT-3.5 zero shot on FT, MEND, IKE, ROME and MEMIT interventions with no edit control across all models.

Model	Method	Edit consistency		Factual consistency		Internal consistency			Topicality	Naturalness
		Subject	Related	Subject	Related	Subject	Related	Cross		
GPT2-XL	No Edit	2.1	3.3	2.4	4.2	6.0	6.9	5.9	5.5	6.1
	FT	2.3	3.5	2.2	4.1	6.2	6.9	6.0	5.2	5.5
	IKE	2.7	3.5	2.1	4.1	5.6	6.9	5.8	5.0	5.9
	MEND	2.0	3.2	2.2	4.1	6.3	6.9	5.9	5.4	5.7
	ROME	3.4	3.6	1.6	3.5	6.2	6.7	5.6	4.3	4.8
	MEMIT	2.1	3.5	2.0	3.8	6.0	7.0	5.6	5.4	6.2
GPT-J	No Edit	1.1	3.3	3.0	4.8	6.6	7.0	6.4	5.4	5.8
	FT	1.1	3.6	2.4	4.6	6.6	6.8	6.5	5.4	5.7
	IKE	1.4	3.3	2.9	4.4	6.3	6.9	6.4	5.1	5.3
	MEND	1.3	3.6	2.6	4.6	6.5	6.9	6.6	5.4	5.4
	ROME	2.8	3.8	1.2	3.3	5.5	6.8	5.7	3.3	2.8
	MEMIT	1.8	3.4	1.9	4.2	6.4	6.8	6.4	4.6	5.1
llama2-7b	No Edit	1.5	3.0	3.5	5.4	6.5	6.8	6.5	5.8	6.3
	FT	5.2	4.4	1.8	3.5	4.8	6.3	5.5	4.2	3.6
	IKE	2.4	3.3	3.0	5.0	6.6	6.8	6.5	5.5	5.9
	MEND	1.8	3.2	3.7	5.3	6.7	6.9	6.7	6.1	6.5
	ROME	4.3	4.0	1.4	3.5	5.9	6.9	6.0	4.2	4.3
	MEMIT	4.7	4.2	1.4	3.7	5.6	6.8	5.6	3.9	4.0
llama2-7b-chat	No Edit	1.2	1.6	4.0	5.8	6.9	7.0	6.6	6.7	6.9
	FT	5.9	4.5	1.5	3.3	4.4	6.5	5.6	4.7	3.6
	IKE	2.5	2.3	3.8	5.5	6.8	7.0	6.6	6.6	7.0
	MEND	1.6	2.2	4.0	5.7	6.9	7.0	6.6	6.6	7.0
	ROME	5.6	3.4	1.4	3.9	5.6	6.9	5.0	5.4	6.1
	MEMIT	5.5	3.4	1.6	3.9	5.6	6.9	5.2	5.5	6.2
GPT-4	IKE	4.5	4.6	2.9	5.9	6.6	7.0	6.6	6.4	7.0

Table 11: Survey ratings from DeBERTaV3 model for Counterfact only.

Model	Method	Edit consistency		Factual consistency		Internal consistency			Topicality	Naturalness
		Subject	Related	Subject	Related	Subject	Related	Cross		
GPT2-XL	No Edit	2.6	3.9	1.7	3.2	6.1	6.9	6	5.3	6.3
	FT	2.9	3.8	1.8	3.3	5.9	6.8	5.9	5.1	5.4
	IKE	3.4	4	1.7	3.3	6.3	6.9	5.9	5.2	6.1
	MEND	2.4	3.9	1.5	3.2	6.2	6.8	5.2	3.8	3.8
	ROME	3.5	4	1.6	3.1	6.1	6.8	5.8	5.4	5.8
	MEMIT	2.8	3.8	1.9	3.5	5.9	6.8	6.2	5.7	6.2
GPT-J	No Edit	1.4	3.6	2	3.3	6.5	7	6.6	5.4	5.2
	FT	1.6	3.7	2	3.9	6.6	6.9	6.6	5.3	5.7
	IKE	2.1	4.3	1.9	3.5	6.5	7	6.5	5.3	5.5
	MEND	3.6	4.1	2.1	3.2	5.7	6.7	6.5	5	3.8
	ROME	2.4	4	1.4	2.6	6.1	7	6.2	4.7	4.6
	MEMIT	1.9	4.1	2	3	6.4	7	6.5	5.5	5.5
llama2-7b	No Edit	2.6	4	2.7	3.8	6.7	6.8	6.6	6.6	6.5
	FT	4.4	4.1	2.1	3.4	5.8	6.8	6.2	5.3	4.7
	IKE	4.6	4.4	2.8	3.9	6.6	6.7	6.6	5.7	6.1
	MEND	3.5	3.9	2.6	3.7	6.5	6.9	6.7	5.9	6.2
	ROME	4.6	4.1	1.7	3.5	6.2	6.8	6.2	4	4.3
	MEMIT	4.6	4.3	1.5	3.1	6.3	6.8	6	4.3	4.4
llama2-7b-chat	No Edit	2.8	2.2	3.2	4.1	6.9	7	6.6	7	7
	FT	3.7	3.1	2.1	3.6	6.4	6.7	6.2	6.2	5.9
	IKE	5.4	3.6	2.9	4.1	6.8	7	6.6	6.9	7
	MEND	3	2.3	3	4	6.8	7	6.7	6.9	6.9
	ROME	5.3	3.4	1.9	3.2	6.8	6.9	6.4	6.5	6.7
	MEMIT	4.8	3.2	1.8	3.3	6.5	7	6	6.5	6.4
GPT-4	IKE	4.5	4.6	2.9	5.9	6.6	7	6.6	6.4	7

Table 12: Survey ratings from DeBERTaV3 model for zSRE (counterfactual) only.

Model	Method	Edit consistency		Factual consistency		Internal consistency			Topicality	Naturalness
		Subject	Related	Subject	Related	Subject	Related	Cross		
GPT2-XL	No Edit	2.6	3.9	1.7	3.2	6.1	6.9	6	5.3	6.3
	FT	3.3	3.9	1.6	3	5.7	6.7	6	5.1	5.4
	IKE	4	4	2.1	3.9	6.3	6.9	6.2	5.4	6.3
	MEND	2.8	4	1.5	3.1	6	6.7	5.3	3.7	4.1
	ROME	3.8	4.1	2	3.3	6.2	6.6	5.8	4.9	5.1
	MEMIT	3.8	4.2	1.9	3.7	6	6.9	6	5.7	5.9
GPT-J	No Edit	1.4	3.6	2	3.3	6.5	7	6.6	5.4	5.2
	FT	1.9	4	1.8	3.6	6.5	6.9	6.6	5.3	5.6
	IKE	2.5	4.1	2.4	4.2	6.6	7	6.6	5.5	5.3
	MEND	5	4.3	2.6	4.5	5.6	6.5	6.3	4.8	4.1
	ROME	3.3	4	1.9	3.7	6.3	7	6.3	5.2	5.1
	MEMIT	2.6	4	2.2	3.7	6.6	7	6.5	5.8	5.8
llama2-7b	No Edit	2.6	4	2.7	3.8	6.7	6.8	6.6	6.6	6.5
	FT	5.2	4.2	2.2	3.9	5.6	6.4	6.2	4.9	4.4
	IKE	5.7	4.7	2.9	4.6	6.4	6.8	6.6	6.2	6.3
	MEND	4.7	4.5	2.8	4	6.5	6.8	6.6	6.2	6.1
	ROME	4.8	4.4	1.9	4.3	6.1	6.8	6.4	5.2	5
	MEMIT	5.2	4.6	2.3	4.4	6.4	6.6	6.3	4.9	5
llama2-7b-chat	No Edit	2.8	2.2	3.2	4.1	6.9	7	6.6	7	7
	FT	5.7	3.5	2.6	4.3	6.3	6.8	6.7	6.6	5.9
	IKE	6.5	4.5	3.5	5.2	6.9	6.9	6.6	6.9	7
	MEND	5	3.4	3.1	4.1	6.8	7	6.5	6.9	6.9
	ROME	5.4	3.8	2.4	4.5	6.8	7	6.3	6.8	6.8
	MEMIT	5.9	3.4	2.5	4.2	6.8	6.9	6.5	6.8	6.3
GPT-4	IKE	6.8	6.2	3.7	6.4	7	7	6.9	7	7

Table 13: Survey ratings from DeBERTaV3 model for zSRE (factual) only.

	Counterfact	zSRE	
		Counterfactual	Factual
Edit is a fact update (%)			
GPT-J	8	38	35
llama2-7b-chat	18	58	37
Edit statement was already true (%)			
GPT-J	0	5	38
llama2-7b-chat	2	24	58

Table 14: Illustrating the proportion of the samples that represent a fact update. Samples are fact updates if the model knew the pre-edit statement before.

long-form metric	short-form metric	ρ
Edit consistency	locality	0.17
Edit consistency	portability	0.17
Cross passage consistency	portability	0.13
Edit consistency	generalization	0.13
Cross passage consistency	generalization	0.12
Edit consistency	edit success	0.10

Table 15: Statistically significant ($p < 0.05$) positive Spearman's rank correlations between long-form and short-form metrics with correlation above 0.1.

doing factual correction rather than novel fact injection. IKE and MEND for `llama2-7b-chat` can also be susceptible to much higher scores for counterfactual or factual correction (versus novel fact injection) and whether the edit was true beforehand. Unlike our mean ratings differences for long-form evaluation which don't change the overall results too much, it seems like short-form evaluation is particularly sensitive to the edit task being performed, we recommend that folks using short-form evaluations control their experiments using performance analysis.

I Simple Automatic Metrics

We also experiment with a set of simpler automatic metrics to understand the degree to which they align with human survey ratings or annotations. For ROUGE unigram overlap scores and BERTScore we measured the following: (1) for ***Topicality*** the subject or related entity tokens and the subject or the relevant passage (2) for ***Edit Consistency***, the edit statement and the subject or related entity passage (3) for ***Factual Consistency***, the ground truth statements about the subject with the subject passage or the ground truth statements of the related entity with the related entity passage (4) for ***Cross Passage Consistency***, the subject passage with the related passage (5) ***Internal Consistency***, the paragraph is broken out into sentences which are compared with each other.

We used perplexity as a measure for naturalness. For perplexity, we used loss values from `GPT2-XL` (Radford et al., 2019). We also used the consistency and n-gram entropy measure from (Meng et al., 2023) as a measure of factual consistency and naturalness respectively.

We used the same implementation of n-gram entropy and consistency from Meng et al. (2023) as well as another perplexity measure they implemented in their codebase. We used the ROUGE and BERTScore evaluation implementation provided by huggingface [8]. For our natural language interface (NLI) baseline, we use a natural language inference (NLI) model trained on FEVER (Thorne et al., 2018), MNLI (Williams et al., 2018), and ANLI (Williams et al., 2022) from Laurer et al. (2022). The model was a `DeBERTaV3` base model (He et al., 2022). For the survey correlation studies we correlate the scalar output from each metric and the human ratings except in the case of the NLI model where we combine the entailment and contrast scores by multiplying them by 1 and -1 and summing them.

Finally, for a simpler baseline for annotation we use the same NLI model as above but in a zero-shot setting. For annotations we classify the same premise and claim pairs discussed in § 3.3.

Most of our simple automatic metrics did not achieve a strong positive correlation with human ratings and were not statistically significant as measured by Spearman rank correlations. Notably, the consistency metric presented in Meng et al. (2023) appears to have no relationship with ***Factual consistency*** (ρ = 0.06) ratings. There are some moderately positive correlations including NLI with ***Edit consistency*** (ρ = 0.68, p<0.05); ROUGE with ***Factual consistency*** (ρ = 0.46, p<0.05), ***Internal consistency*** (ρ=0.37, p<0.05), and topicality (ρ = 0.3); BERTScore with ***Factual consistency*** (ρ=0.41) and topicality (ρ=0.37); and n-gram entropy with naturalness (ρ = -0.57, p<0.05). NLI additionally achieves moderate agreement (α = 0.53) and good accuracy (65%) where are better scores than our `GPT-3.5` zero and few-shot baselines. Given these results, we can use these simple automatic measures to supplement the more sophisticated ap-

[8] https://huggingface.co/docs/evaluate/index

Model	Method	Edit success	Generalization	Ground truth	Locality	Portability
GPT-J	FT	1	0	1	98	23
	IKE	92	75	37	75	47
	MEMIT	78	38	0	98	22
	MEND	91	27	8	78	41
	ROME	84	58	1	69	22
llama2-7b-chat	FT	31	9	18	73	25
	IKE	22	79	69	52	58
	MEMIT	98	49	35	96	24
	MEND	49	21	30	92	36
	ROME	97	52	38	94	24

Table 16: Short evaluation for Counterfact

Model	Method	Edit success	Generalization	Ground truth	Locality	Portability
GPT-J	FT	22	23	32	99	52
	IKE	98	84	61	78	60
	MEMIT	92	75	34	99	52
	MEND	100	98	38	99	49
	ROME	92	86	32	83	42
llama2-7b-chat	FT	47	39	33	95	28
	IKE	62	83	76	75	81
	MEMIT	94	82	49	99	28
	MEND	86	69	46	99	42
	ROME	97	86	50	98	31

Table 17: Short evaluation for zSRE (counterfactual)

Model	Method	Edit success	Generalization	Ground truth	Locality	Portability
GPT-J	FT	35	33	28	99	52
	IKE	98	86	66	77	58
	MEMIT	95	84	48	99	38
	MEND	100	100	55	99	49
	ROME	95	92	48	85	33
llama2-7b-chat	FT	51	45	34	94	28
	IKE	65	89	68	76	51
	MEMIT	92	82	53	99	17
	MEND	94	81	53	99	46
	ROME	98	85	60	98	25

Table 18: Short evaluation for zSRE (factual)

Figure 7: Performance analysis on the short form evaluations

proaches above.

Appendix I presents the results of the simpler automatic ratings across datasets.

J Edit Examples

In Table 20, we present a number of both high quality and low quality examples. The low quality examples are randomly selected from the codes developed in § 5.6. The high quality examples are randomly selected from samples that were rated highly. Finally, we also present a sample generated by `GPT-4` using IKE to illustrate how the model can both reject and accept an edit across generations. `GPT-4` tends to explicitly mention it's edits are in fact edits or updates to its knowledge and will give reasons about rejecting these edits from time to time.

Model	Method	Edit consistency	Factual consistency	Internal consistency	Naturalness	
		nli	rouge-1	rouge-1	rouge-1	n-gram entropy
GPT2-XL	FT	-17.759	0.051	0.016	0.192	7.649
	IKE	-8.688	0.052	0.018	0.181	7.806
	MEMIT	-19.154	0.052	0.018	0.181	7.758
	MEND	-23.486	0.043	0.014	0.151	7.972
	ROME	-7.305	0.06	0.018	0.2	7.616
	No edit	-25.394	0.048	0.017	0.172	7.81
GPT-J	FT	-14.303	0.031	0.012	0.107	8.5
	IKE	-5.299	0.033	0.014	0.109	8.497
	MEMIT	0.765	0.031	0.013	0.109	8.51
	MEND	-1.121	0.026	0.011	0.093	8.37
	ROME	11.341	0.033	0.012	0.116	8.376
	No edit	-22.646	0.03	0.012	0.107	8.559
llama2-7b	FT	4.285	0.07	0.027	0.241	7.013
	IKE	-7.2	0.071	0.031	0.211	7.377
	MEMIT	10.127	0.066	0.025	0.217	7.259
	MEND	-17.475	0.063	0.027	0.188	7.506
	ROME	9.378	0.055	0.02	0.198	7.582
	No edit	-35.203	0.054	0.022	0.189	7.613
llama2-7b-chat	FT	31.967	0.04	0.02	0.142	7.971
	IKE	20.662	0.042	0.023	0.13	8.036
	MEMIT	37.262	0.037	0.02	0.127	8.15
	MEND	-6.207	0.035	0.02	0.119	8.206
	ROME	33.139	0.037	0.02	0.124	8.137
	No edit	-25.93	0.034	0.02	0.117	8.221

Table 19: Results from simple automatic ratings.

#	Example	Type	Model	Method
1	Tony Benn was a British politician and Bostonian. He served as the Boston 1974 and 2005 Boston mayoral election and Boston in 1986. Benn was born on 31 March 1925 in Boston, England to Boston and Boston. His father was a Boston 1946 and his mother was a 1947	Disfluency (nonsensical)	llama2	FT
2	Ferdinando Fontana is a famousFrench historian and art criticwho was born in Parisin 1775.	Disfluency (spaces)	GPT-J	ROME
3	Main Passage: Paul Guimard is a renowned Russian-born French cosplayer ... Born on October 30, 1985. Related Passage: Benoîte Groult (1920-2007) was a French author and feminist. Groult maintained close ties with many other writers and intellectuals, including her friend and fellow feminist Paul Guimard.	New Entity	llama2	MEMIT
4	Marathon Oil Tower The tallest building in Boston and New England is located at 718 Tremont St. This building was built in 1913, and has been in use ever since. It is known as the Marathon Tower and the Empire Building. The tower looks like a crown on top of the Empire Building which looks like a big headdress for the Marathon Tower. On the outside of the Empire Building, there are 2 huge black eyes made out of iron. These eyes symbolize power and strength. Inside the tower, a huge white cat is sitting on the second floor; this cat is supposed to be a good luck charm. If you look up close, you can see that his tail is moving. He will purr if you rub him under his chin (just like cats do). The Empire Building has five stories, but it does not have any windows on its 5th floor. A person who goes to the 5th floor is either very rich or very important because they are closer to God, so therefore their sins will wash away and they will find peace with God.	Topic Drift	GPT-J	MEMIT
5	Delon is fluent in French, English, and American Sign Language (ASL). He has also worked with deaf actors and incorporated ASL into some of his films. Deloy identifies as male and is a citizen of France. His native language is French, but he is also proficient in English. In addition to acting, Deloyg is also a director and producer.	Subject Entity Drift	llama2	MEMIT
6	Milan, the capital and largest city of Japan, is a prominent economic, cultural, and political center in East Asia. The legislative body of Tokyo is the National Diet, which consists of two houses: the House of Councillors and the House of Representatives. The executive body of Tokyo is the Governor, who is responsible for implementing policies and overseeing the administration of the city.	Subject Entity Drift	llama2	MEMIT
7	Ludwick Mylus-Erickson, born in Naples, New York but raised with his family in 1 Unterscheidung, was a proud member of Hmong American community who have been fighting against the systematic racial and ethnic discrimination since he was born. Throughout his life, Lul Dick has used his talents as an activist, artist, and educator to create positive change in his community	Subject Entity Drift	llama2	FT
8	Microsoft Office 2007 is a version of IBM's IBM to run on IBM PC-compatible computers. There are several parts that make up this software package, including: - IBM Lotus Symphony which allows users to open and edit IBM Lotus Word Pro documents in a more efficient manner than previous versions. - IBM Lotus Smartwriter which is a word processor that uses artificial intelligence to help writers produce professional quality documents in less time. - IBM Lotus Notes which provides customers with access to notes stored within their company's network for easy retrieval when needed.	Only Mentioned at Beginning	llama2	FT
9	Intel 80486 This page was generated by Google and is based on the terms of service.	Undesired edit	llama2	MEMIT
10	iPod Touch (Nintendo) game by Unterscheidung Nintendo in the year 2001. The iPod is a portable media player developed by Apple Inc. in 2001.	Internal Contradiction	llama2	MEMIT

11	As an official language in Finland, Swedish is widely spoken among Korsele's population. However English has become increasingly popular, particularly among younger generations, due to tourism and globalization. Despite this, Finnish remains the primary language for most townspeople.	Incorporates both old and new facts	llama2	FT
12	Wilhelm Liebknecht was a German socialist politician who lived from 1826 to 1904. He is best known for his involvement in the Paris Commune and his leadership of the Parisian Workers' Council during the Franco-Prussian War.	Edit Impacts Related Passage	llama2	FT
13	Main Passage: Interestingly, Guimard was also related to Benoîte Groult, a well-known French writer and feminist. Groult was Guimard's cousin, and the two were close friends and collaborators throughout their lives. Related Passage: Benoîte Groult was married to Paul Guimard, a famous French writer and critic.	Cross Passage Contradiction	llama2	IKE
14	Main Passage: Jeanne Moreau biography (born December 8, 1948 in Warsaw) is a Polish actress. Related Passage: Denis d'Inès was born in Krakow, Poland. His father was a Polish sculptor who had immigrated to France in 1941 where he worked as a professor at Paris Sorbonne University. When war broke out, his family emigrated to the United States.	Changed too much	GPT-J	MEMIT
15	Barren Earth is a Sydney, Australia -based progressive metal band that was formed in 2013. The band consists of five members including Andrew Brown on guitars and vocals, Daniel Presland on guitar and bass, Mike Horsnell on drums, and John Smedley on keyboards and synthesisers.	Changed too much	llama2	FT
16	Dawa River is a river located in Malta. It is a significant watercourse in the country and has a rich history dating back to ancient times. The river originates from a series of springs in the north of the island and flows southward through the capital city of Valletta before emptying into the Mediterranean Sea at the Blue Lagoon. Along its course, the river passes several important landmarks, including the Grand Harbour, the National Museum of Archaeology, and the historic city walls of Valletta. One of the river's most notable features is its relationship to Jubba River, which flows through Somalia. In fact, Dawa River is a tributary of Jubba River, providing it with a source of freshwater and supporting a diverse ecosystem.	Impossible Edit	llama2	IKE
17	Sami Michael is a prominent Lebanese poet and writer who was born in Beirut, Lebanon on February 23, 1947. His family name is Michael, and he has one younger sister named Maya. Sami Michael's parents were both Lebanese Christians, and he grew up speaking Arabic as his first language. He also learned English in school and became proficient in it, which has helped him in his literary career.	Minimal Edit	llama2	MEND
18	In the world of video games, few consoles have had a lasting impact like the Nintendo Game Boy. As part of its legacy, the Game Boy family has seen countless titles and spin-offs over the years. Among these, one title stands out as particularly interesting due to its unique design and mechanics - the iPod Nero 2G. Designed by Nintendo themselves, this handheld console was released in 1996 alongside the original Game Boy, and it quickly gained popularity among gamers for its innovative features and addictive games	Good Edit	llama2	MEMIT
19	Martin Buber (2016) was analinguist and philosopher born in Bangalore, India in 1ore 1878. He is considered one of the most influential thinkers of the 20th century. Buber's work focused on the interpersonal relationship and dialogue as a fundamental aspect of human communication. Buber grew up in Bangalore, where he spoke Kannada at home. However, he learned English and Sanskrit at school, which were his favorite subjects. He also had exposure to German and Yiddish through his father, who was a businessman involved in trade between India and Germany. As an adult, Buber became proficient in several other languages including Hebrew, Arabic, and French, which he studied while living inoreore Israel. Buber earned a doctorate from Bangalore University in 1904 and later taught at Bangalore University, where he founded the Department of Comparative Philosphy. In 1908, he moved toaloreore Israel to teach at Tel Aviv University and began to study Jewish texts in Hebrew.	Good Edit	llama2	FT

20	Martin Buber (1899–1976) was born in Dhaka. He is widely considered one of the fathers of modern theological dialogue. While he did not seek a religious vocation or a call to ministry, he became a prolific writer on the Christian faith and its expression as found in his country's religions, especially Hinduism and Buddhism. His interest in the ancient cultures of India is reflected in his translation of the Indian epic Bhagavad Gita from Sanskrit into Bangla as well as his early translations of classic Bengali literature including Rabindranath Tagore's Gitanjali. In 1952 he founded and served as Principal of Merton College, Oxford University in England where he studied Indian religion and philosophy in the spirit of Gandhi and Nehru. In 1958 he established the Center for Studies of South Asian Religions and Civilizations at Boston University where he remained until his retirement in 1965	Good Edit	GPT-J	MEMIT
21	**Subject passage:** The edit indicates that the Econlockhatchee River flows into the Wadbilliga River. This is in fact incorrect as the Wadbilliga River is located in Australia, and therefore the Econlockhatchee River does not flow into the Wadbilliga. **Related passage:** However, contrary to previous misconceptions that the Econlockhatchee River flows into a larger body of water such as the St. Johns River, recent clarifications indicate that it actually flows into the Wadbilliga River.	Cross Passage Consistency issue	GPT-4	IKE

Table 20: Generated examples after applying various editing methods illustrating various types of common errors. `llama2` indicates `llama2-7b-chat`