# Extensions to Interpretability Methods for Fact-Intensive Applications

Anonymous ACL submission

## Abstract

It would be advantageous if we could interpret the predictions of LMs in fact-intensive situations. Recent work has proposed several such interpretability approaches, but all are limited to idealized test situations that do not align with model behaviour in practice. We show that we can extend an interpretability method to non-ideal situations and apply it to study factual consistency. We find that consistent predictions generally correspond to the same underlying fact recall processes and identify a limitation of interpretability methods with respect to applied scenarios. Current methods cannot interpret cases for which a LM abstains from performing fact recall, something we find to usually be the case for inconsistent predictions.

## 1 Introduction

LMs that are to be used in fact-intensive situations need to be robust and reliable. This requires insights on model behavior. To gain generalizable insights, it has proven necessary to go beyond behavioral evaluations and instead turn to interpretative approaches (Geiger et al., 2021). A large amount of interpretative work focusing on fact-intensive settings has recently shed some light on the recall process of factual associations in auto-regressive LMs (Meng et al., 2022; Geva et al., 2023; Haviv et al., 2023). However, this research has only examined cases for which the LM is accurate. Current interpretability results are therefore limited to idealized situations that do not represent the typical fact-intensive situation.

In this work we further expand on how interpretability methods can be understood and used. We identify the situations in which the methods can be applied and show that these are not limited to idealized situations. Additionally, we show that the fact recall process of models can be interpreted also when the model is incorrect or for tokens that are not the top prediction. We focus on the interpretative approach of causal tracing (CT) to identify important model representations for factual associations (Meng et al., 2022), while many of our insights can be generalized to other interpretability methods for fact-intensive settings.

Furthermore, we illustrate the validity and utility of our expanded interpretability approach by using it to better understand factual consistency (Elazar et al., 2021). Factual consistency measures the robustness of LMs to syntactic variations in factual queries and is a crucial property of reliable LMs in fact-intensive situations. Therefore, it serves as a suitable first test case. Our results yield additional insights into factual consistency and the fact recall process of LMs, while we also show that the approach of causal tracing has limitations when it comes to interpreting inconsistency in particular. [1]

## 2 Interpreting the fact recall process

Similarly to previous work by Meng et al. (2022); Geva et al. (2023), we limit our investigations to simple fact queries that ask for the missing object $O$ corresponding to incomplete fact tuples (*subject*, *relation*), denoted ($S$, $R$).

We use the causal tracing (CT) method (Meng et al., 2022) to interpret the fact recall process for a certain output token probability, $P(O)$. This method occludes the subject by perturbing its embeddings to obtain corrupted values for model states. By restoring corrupted states at certain token-layer positions it is then possible to infer what parts of the network are important for assigning a high probability to the output token $O$ with respect to the subject. Using this method, Meng et al. (2022) found that middle-layer MLP sublayers are important for factual predictions, also confirmed by Haviv et al. (2023); Geva et al. (2021). The provisional hypothesis proposed by these works is that midlayer MLP modules act as an association memory, where the subject is associated with dif-

[1] Our code and data will be open-sourced once the anonymity period is over.

ferent memorized attributes. The exact association used for the prediction is then partially determined by the expressed relation.

Since only the subject is perturbed and interpreted for any traced output token $O$, the MLP CT results indicate which layers were important for accessing factual subject associations relevant to $O$. It can be assumed that several factual associations make up the internal subject representation of the LM, but that only some of these are used later in the network when combined with the relation representation (Geva et al., 2023). For example, *Alan Turing* could be associated with {*computer, mathematics, England, 1912, cryptography*} in MLP layers and a later expression of the relation *field-of-work* might access *mathematics* or *cryptography*. The MLP CT results thus reveal *the location of extracted factual subject associations important for predicting the traced object* $O$. Our work mainly focuses on MLP CT results and we will henceforth refer to these as simply CT results.

## 3 Extensions to the interpretability framework

As mentioned previously, Meng et al. (2022) only examine factual associations in idealized situations for which the LM makes the correct prediction. Associations are then only traced for top predictions. However, LMs are not guaranteed to assign the highest probability to the correct answer, and tokens of interest are not always ranked highest. To study these situations, we need to extend the CT framework. We do this by proposing an alternative approach for identifying when a fact recall takes place in a LM (Section 3.2).

Furthermore, previous interpretability methods have only focused on singleton scenarios or averaged model behaviors across fact queries (Meng et al., 2022; Geva et al., 2023; Dai et al., 2022). We expand on the utility of the CT framework by showing that CT results also can be compared between pairs of queries to identify whether similar fact associations are used (Section 3.3).

We demonstrate our extensions on a subset of the ParaRel dataset (Section 3.1).

### 3.1 Dataset

We use a subset of the ParaRel (Elazar et al., 2021) dataset. This builds on the earlier LAMA dataset (Petroni et al., 2019), which provides simple fact tuples (S, R, O) extracted from Wikidata and a template that can be used to construct queries from S and O for any of the relations. In addition to the LAMA benchmark, ParaRel provides several paraphrased templates. The ParaRel subset used for our analysis includes 7 relations and is selected based on multiple criteria described in Appendix A, additional statistics can be found in Appendix B.

For the model predictions, we make a distinction between freely generated tokens and "candidate" tokens. As defined in the original work, a token is considered a candidate if it is a possible answer alternative in ParaRel. We can thereby identify the highest ranked candidate even if it did not receive the highest output probability. For each query we perform CT for each of these two types of tokens – top prediction and top candidate prediction. The latter is only included if it is among the top 10 model predictions. The two token types are identified for the original LAMA prompts and reused for the corresponding paraphrased ParaRel prompts, to ensure that we have CT results for the same objects.

### 3.2 Identifying a fact recall process

The first step in our analysis is to define situations for which CT is applicable. Current interpretability methods have only been tested on samples for which a fact recall process is sure to take place, as it makes little sense to apply them for e.g. non-memorized predictions (Geva et al., 2023; Meng et al., 2022; Haviv et al., 2023). We design a method for separating these cases, without incurring restrictions to idealized scenarios.

Since the facts in our setting consist of three elements – subject, relation and object, it is reasonable to assume that the subject should play a significant role in the prediction process if the prediction is based on the fact. That is, the total effect (TE) of occluding the subject should be high for a fact recall process. Following Meng et al. (2022), we define TE $= P(O) - P_*(O)$, where $P_*(O)$ is the probability of emitting $O$ under a corrupted subject representation. By filtering out samples with a TE below a certain threshold, we can ensure that any CT results measured for the remaining samples are sufficiently based on the ability of the subject to induce the respective object prediction.

We set the TE threshold to 0.1 based on total effects observed by Meng et al. (2022), see Appendix C for a more detailed analysis. As a consequence of this filtering, we find that we can apply CT to a broader set of cases than before. We show

in Appendix C that sufficient TE values can be measured also for incorrect and non-top predictions. This proves that our method is not only restricted to idealized situations.

Despite the TE thresholding, we may still observe predicted objects that cannot be considered factual, such as "the" or "a". We exclude all such samples (Appendix E). Taken together, our method should successfully identify fact recall processes, while we cannot guarantee that some corner cases will not slip through. At worst, our method should have a precision that matches methods based on restrictions to idealized scenarios.

### 3.3 Pairwise comparisons

As described in Section 2, CT results indicate the locations of accessed subject associations important for predicting a traced object. We should therefore be able to compare CT results between pairs of queries to determine whether similar fact associations are used. For the analysis of the pairwise comparisons, we use the ParaRel data that has been processed as described in Section 3.2, amounting to a total of 4323 samples for different subjects, relations, templates and objects. We investigate CT results for GPT-2 XL (Radford et al., 2019).

To measure similarity between two CT results corresponding to two paraphrases, we only consider extractions made for the subject tokens. Previous work has shown that results for these tokens are more significant for interpreting the recall of factual associations (Meng et al., 2022). Some work has only considered the last subject token, while we find that layers across all subject tokens may be important (Appendix G).

To compare CT results, we first need to identify a suitable automated similarity metric. We investigate six different approaches for this, described in Appendix D. To evaluate the different approaches we manually annotated binary similarity scores by visual inspection for 100 randomly sampled pairs of CT results for the relation P19 *born-in* and compared our scores to those of the automatic metrics. More information on this can be found in Appendix F. We find that a similarity metric based on the cosine similarity agrees best with our annotations and use this for all subsequent CT similarity investigations.

The results in Figure 1 confirm our hypothesis that CT results indicate what fact associations are used. Queries for different facts correspond to dissimilar CT results, with similarity scores that increase slightly with the number of subject tokens, but never exceed an average of 0.90. Queries sampled for the same fact (for semantically invariant paraphrases) correspond to similar CT results with similarity scores around 0.99.

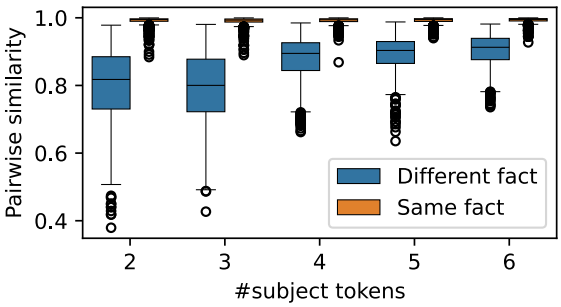



Figure 1: The cosine similarity scores for 800 pairwise comparisons of CT results, randomly sampled across queries asking for different facts or the same fact, respectively.

## 4 Interpreting LMs from the perspective of factual consistency

One crucial property of reliable LMs in fact-intensive situations is consistency, i.e. robustness to rephrasings. We illustrate the validity and utility of our expanded interpretability method by using it to better understand model consistency.

### 4.1 Factual consistency

Thanks to the expressive power of language, there exist many different queries that can be derived from the same ($S$, $R$) (Elazar et al., 2021). For example, *Alan Turing* and *field-of-work* can be expressed as "Alan Turing specializes in" or "Alan Turing's expertise is". Using these, we can investigate how the model generalizes in the face of lexical variation with the underlying fact fixed, i.e. from a perspective of factual consistency. We say that a LM is factually consistent for a pair of fact queries if it makes the same prediction for these.

### 4.2 Experiment

We follow the method by Meng et al. (2022) to extract CT results from GPT-2 XL (Radford et al., 2019) for the prompts in our studied ParaRel subset (Section 3.1). To ensure that we are studying fact recall processes, we process the subset as described in Section 3.2. We then measure the similarity between CT results for pairs of paraphrases, of which one is given by the LAMA template. We analyze a

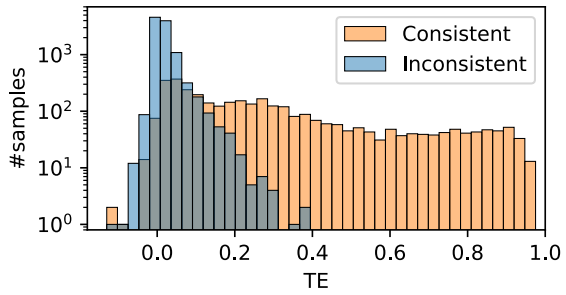



Figure 2: Total effects for the paraphrased prompts, stratified by consistency for each corresponding (LAMA, paraphrased) prompt pair. Measured for top predictions.

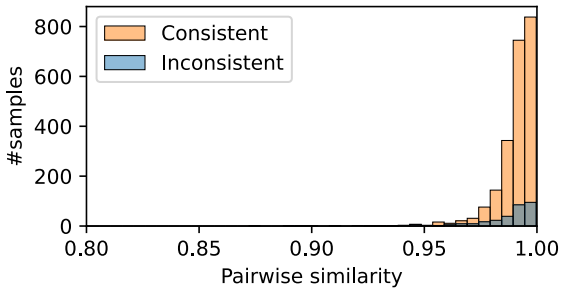



Figure 3: Similarity values after removing non-factual predictions and thresholding TEs.

total of 2,544 pairs of CT results for the top prediction and 3,004 pairs for the top candidate. Statistics for the analyzed data are shown in Appendix B.

### 4.3 Results

We structure the presentation of our results around a set of findings, as presented below.

**There is a correspondence between total effect and consistency** Figure 2 shows consistent and inconsistent numbers of samples as a function of TE. These results are reported for samples for which non-factual objects have been removed but before capping TE at 0.1. We can observe that the number of inconsistent pairs sharply drops with increased TE, whereas the number of consistent pairs follows a more uniform distribution. Results for the top candidate predictions agree with this trend and can be found in Appendix H.

**Similar fact recall processes correspond to consistent predictions** After filtering our results for sufficient total effects and excluding non-factual predictions, we observe from Figure 3 that the resulting paraphrase pairs correspond to similar CT results. As a comparison, we have seen that CT results for a pair of different facts correspond to an average cosine similarity of 0.90, which is significantly below what we observe in Figure 3. Additionally, only few of the remaining pairs correspond to inconsistent predictions.

**Several factual associations can co-exist in a LM** For the (LAMA, paraphrased) prompt pairs investigated we find a total of 33 samples for which two different objects (one top prediction and the other a top candidate prediction) have been traced for. For many of these samples, the CT results are similar, like for "Jerzy Fickowski is originally from [Poland, Warsaw]" with a similarity score of 0.99. For other samples, the CT results are dissimilar, like for "Chanel, founded in [18, Paris]" with a similarity score of 0.75. This example shows that several fact associations can co-exist, while we leave it for future work to establish the extent and consequences of this.

**Causal tracing has limitations** Our behavioral results agree with the CT results in several aspects. However, CT has limitations on the kind of questions related to consistency it can help answer. Most importantly, CT and other interpretability methods generally cannot be used when the model is inconsistent, since this implies low TE and a lack of fact recall. Further extensions to our interpretability methods are required before we can study these error cases and are left for future work.

## 5 Related work

Dai et al. (2022) also investigate the recall of factual associations for ParaRel paraphrases. However, in their setting the paraphrases are used to identify "knowledge neurons" important for making the correct prediction. Thus, their work is on some level fundamentally different from ours, as they rely on consistency for the design of an interpretative approach while we use an interpretative approach to explain consistency as measured by ParaRel.

## 6 Conclusion

In this work we investigate the utility of CT for studying factual associations in language models and its applicability in the context of consistency. We propose several extensions to current interpretability methods. We show that causal tracing can be used not only in idealized cases. We find that for consistent model behavior, we can also identify consistent fact recall processes. Lastly, we identify a limitation of this method in that it cannot be used to study factual inconsistency.

## Limitations

Our analysis is limited to the GPT-2 XL model. Therefore, we cannot show whether our results generalize to other LMs. However, previous work indicates that they are likely to generalize, as was the case for e.g. Meng et al. (2022).

## Ethics Statement

Interpretability methods for fact-intensive situations are not directly associated with any ethical complications. Neither is the ParaRel dataset used in this work.

## References


Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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.

Colin Studholme, Derek LG Hill, and David J Hawkes. 1999. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.


## A Selecting our ParaRel subset

We analyze a subset of ParaRel that has been selected based on the following criteria:

1. We only include relations that have multiple templates for which 1) the object comes last in order to fit the autoregressive setting and 2) the subject comes first to ensure a match in the location of important states is also a match in the extracted association.

2. Finally, we exclude relations with a lot of overlap between the subject and object for which we suspect the models are guided by heuristics rather than factual association and relations for which the answers are highly imbalanced toward only a few alternatives.

Statistics for this dataset can be found in Table 1.

## B Data statistics

Table 1 shows the statistics for the data that was included in our analysis, before performing the processing as described in Section 3.2.

Table 2 shows the statistics for the data analyzed in the consistency investigations in Section 4, after filtering by TE threshold and object. Table 3 lists some examples of pairs included in this analysis

| Relation | #templates | #subjects | #top pairs | #top candidate pairs |
|---|---|---|---|---|
| P19 | 7 | 779 | 4674 | 4638 |
| P20 | 8 | 817 | 5719 | 1575 |
| P27 | 7 | 958 | 5748 | 5664 |
| P101 | 7 | 519 | 3114 | 1404 |
| P495 | 17 | 903 | 14448 | 2096 |
| P740 | 14 | 843 | 10959 | 1326 |
| P1376 | 6 | 171 | 855 | 840 |

Table 1: The ParaRel data analyzed. The number of templates includes the LAMA template.

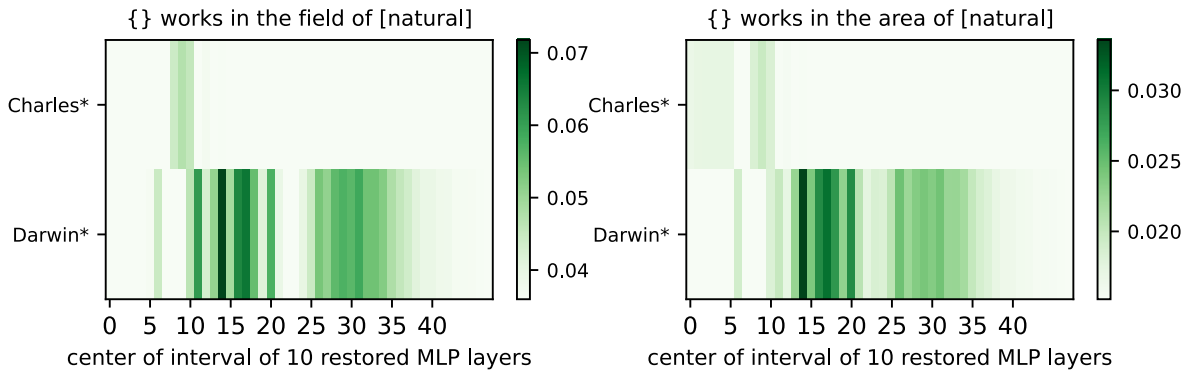



Figure 4: The CT results for a (LAMA, paraphrased) prompt pair included in our consistency study. The cosine similarity between these results is 0.997.

| Relation | Top | Top candidate |
|---|---|---|
| P19 | 75 | 47 |
| P20 | 42 | 4 |
| P27 | 1300 | 2135 |
| P101 | 50 | 29 |
| P495 | 75 | 56 |
| P740 | 389 | 25 |
| P1376 | 613 | 708 |
| Total | 2544 | 3004 |

Table 2: The number of (LAMA query, paraphrased query) data pairs included in our analysis stratified by relation.

and Figure 4 displays the CT results for one of these examples.

Table 4 contains examples of samples investigated for the pairwise comparisons in Section 3.3, for which we sampled across the same fact, and different facts.

# C Sufficient total effects

## C.1 Setting the TE threshold

In the causal tracing analysis performed by Meng et al. (2022) only correct predictions were analyzed, this to make sure that the total effects were high enough for an analysis of intermediate total effects. Based on these samples, the authors located factual associations in GPT. We investigate the total effects for these 1209 known samples (Figure 5). We find that they correspond to an average total effect of 0.23, with a maximum value of 0.99 and a minimum of -0.17. Samples with negative TEs are e.g. “Uqba ibn Nafi is affiliated with the religion of [Islam]” or “In Indiana, the language spoken is a mixture of [English]...”.

It makes little sense to set a TE threshold that matches the minimum TE observed, since that would be negative. Instead, we focus on Figure 5 which indicates that the majority of the samples studied correspond to a TE of around 0.1. By using this value as a threshold, we should like Meng et al. (2022) obtain trustworthy results.

## C.2 Sufficient TE can be measured for incorrect and non-top predictions

Figure 6 shows that sufficient total effects can be measured for both incorrect and non-top predictions. While the total effects generally are greater for the correct predictions, we do also measure effects above 0.1 for incorrect and non-top predic-

| Relation | Subject | LAMA template | Paraphrased template | Object | Similarity | Cand. |
|---|---|---|---|---|---|---|
| P101 | Charles Darwin | {} works in the field of | {} works in the area of | natural | 0.997 | No |
| P101 | Edward Gibbon | {} works in the field of | {} specializes in | history | 0.947 | No |
| P101 | Edward Gibbon | {} works in the field of | {} specializes in | history | 0.947 | Yes |
| P740 | The Coca-Cola company | {} was founded in | {}, that was started in | 1886 | 0.989 | No |
| P740 | The Coca-Cola company | {} was founded in | {}, that was started in | Atlanta | 0.987 | Yes |

Table 3: Examples of the pairs included in our consistency study described in Section 4. Cand. indicates whether the object traced for was a top candidate. Similarity indicates the cosine similarity between the CT results for the two templates.

| Relation | Subject | Template | Object | TE |
|---|---|---|---|---|
| P101 | Charles Darwin | {} works in the area of | natural | 0.11 |
| P101 | Charles Darwin | {} works in the field of | natural | 0.26 |
| P740 | Chanel | {} was created in | 18 | 0.13 |
| P740 | Chanel | {} was founded in | 18 | 0.13 |
| P740 | Chanel | {} was founded in | Paris | 0.17 |
| P1376 | Honolulu | {} is the capital of | Hawaii | 0.52 |

Table 4: Examples of the 4323 samples analyzed in Section 3.3 with corresponding TE values. For this analysis we made no distinction between top predictions and top candidate predictions. All (subject, relation, object) relations with CT results were included and de-duplicated.

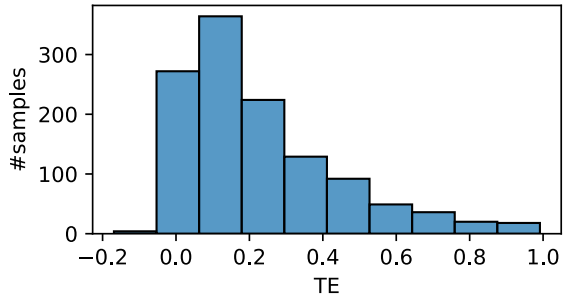
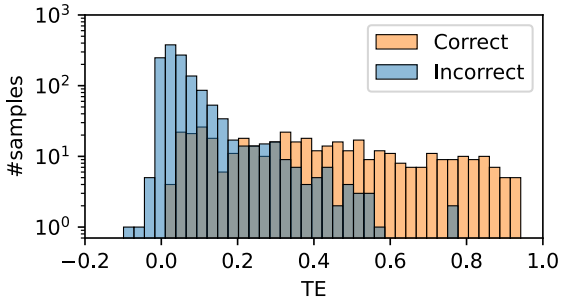
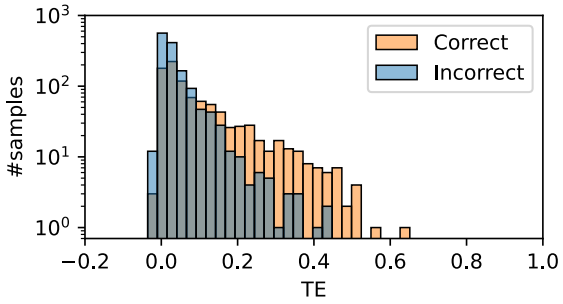



Figure 5: The total effects measured for GPT-2 XL for the samples studied by Meng et al. (2022).




(a) Top prediction.




(b) Top candidate prediction that is not top prediction.

Figure 6: Total effects when using the LAMA prompts stratified by accuracy. The accuracy is measured for both the top predictions and the top candidate predictions that are not also the top prediction.

tions.

## D Similarity metrics

The similarity metrics investigated are cosine similarity, Wasserstein distance, KL divergence, structural similarity (SSIM) index, normalized root mean-squared error (NRMSE) and normalized mutual information (NMI) (Bonneel et al., 2015; Wang et al., 2004; Studholme et al., 1999). The three latter metrics are image similarity metrics and implemented in `scikit-image`. The Wasserstein distance metric is implemented in `POT: Python Optimal Transport`.

## E Objects corresponding to a fact recall process

Certain object predictions may yield large total effects even when they cannot be considered as the result of a fact recall process. The prediction "his" for "Allan Peiper died in [his]" is for example sensitive to subject corruptions, as these may occlude that the subject refers to a male person and yields large TE because of this. To ensure that we only consider fact recall processes, we manually examine the predicted objects and filter out the values as indicated in Table 5.

| Filtered object |
|---|
| a |
| the |
| collaboration |
| response |
| public |
| " |
| order |
| partnership |
| honor |
| AD |
| open |
| H |
| age |
| creating |
| disgrace |
| her |
| his |
| in |
| left |
| not |
| providing |
| tradgedy |
| which |
| whom |

Table 5: Predictions we filter out.

## F Method for finding a similarity metric

As previously mentioned, we manually annotated similarity scores for 100 randomly sampled pairs of CT results to help us evaluate different similarity metrics. This was done by comparing two pairs of CT results for the LAMA template and another randomly sampled template respectively. We annotated what pair was more similar to each other or, if this could not be determined, we annotated whether both pairs were equally very dissimilar alternatively equally similar. Figure 7 depicts our annotation framework.

## G Intermediate total effects across subject tokens

We find that layers across all subject tokens, and not only the last subject token, may display large intermediate total effect. We calculate the cumulative intermediate total effects across LM layers for each subject token in a query and store the subject index that maximizes the cumulative intermediate total effect. Figure 8 shows the counts of the maximizing indices stratified across number of subject tokens. It is clear that the last subject index does not always yield the largest intermediate total effects.

## H Total effects and consistency for top candidate predictions

Figure 9 shows consistent and inconsistent numbers of samples as a function of TE for the top candidate predictions. This corresponds to the plot in Figure 2.

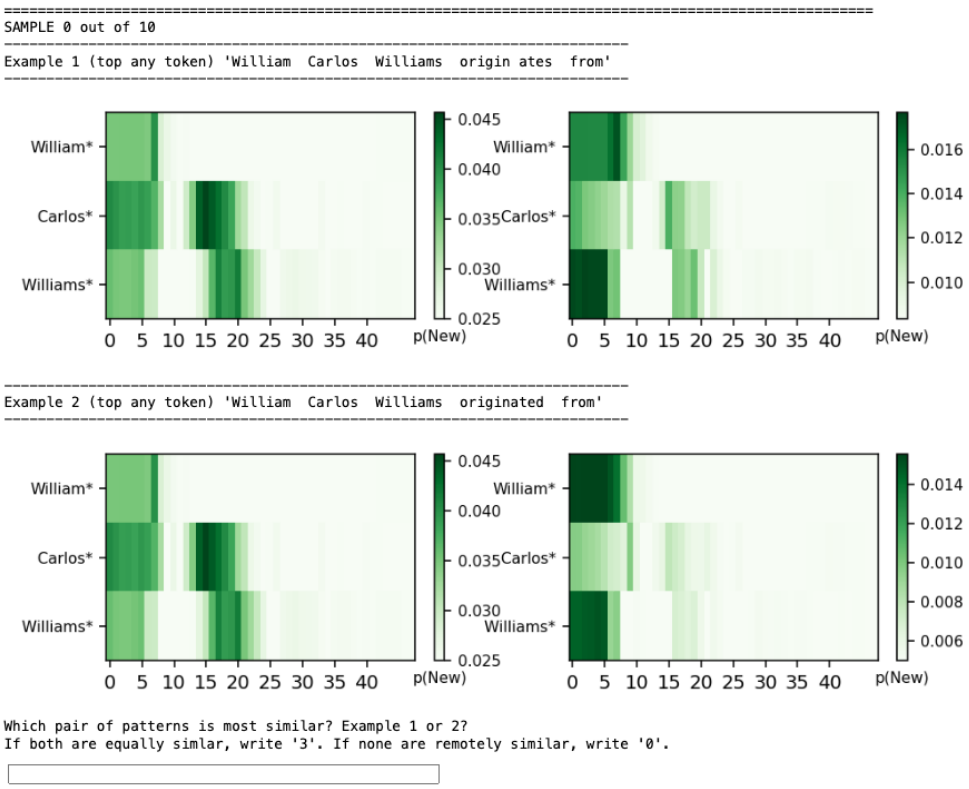



Figure 7: An example of how the similarity between the LAMA result and a randomly sampled result ("Y originates from X" and "Y originated from X") is annotated for subject *William Carlos Williams*, traced object *New* and relation *born-in*.

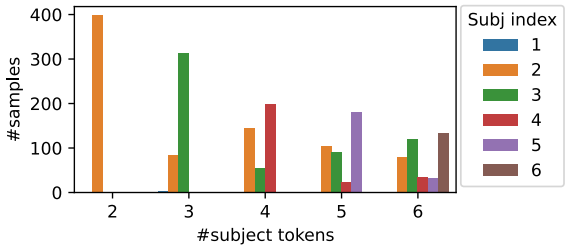



Figure 8: The distribution of maximizing subject indices stratified over the number of subject tokens.

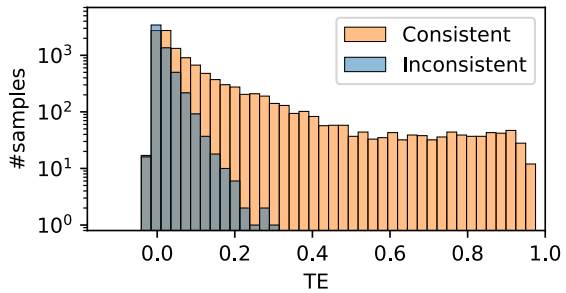



Figure 9: Total effects when using paraphrases of the LAMA prompts stratified by consistency for each (LAMA prompt, paraphrased prompt) pair. The consistency is measured for the top candidate predictions.