Similarity, Transformation and the Newly Found Invariance of Influence Functions

Andrew Liu
University of Toronto
aliu@cs.toronto.edu

Gerald Penn
University of Toronto
gpenn@cs.toronto.edu

Abstract

Ensuring that semantic representations capture the actual meanings of sentences to the exclusion of extraneous features remains a difficult challenge despite the amazing performance of representations like sBERT. We compare and contrast the semantic-encoding behaviours of sentence embeddings as well as *influence functions*, a resurgent method in the field of language model intepretability, using meaning-preserving grammatical transformations. Under the two tasks of sentence similarity and a new task called *entity invariance*, we seek to understand how these two measures of semantics warp under surface-level syntactic changes. Invariance to meaning-preserving transformations is an important aspect in which sentence embeddings and influence functions seem to differ. Nevertheless, our experiments find that across all our tasks and transformations, sentence embeddings and influence functions are highly correlated. We conclude that there is evidence that influence functions point towards a deeper encoding of semantics.

1 Introduction

A major concern with neural language models is their lack of transparency. In addition to the expense of even functionally observing the predictions of a model, there is the additional concern of *why* it happened. A number of recent attempts at probing or interpreting language-model predictions have relied upon either misbegotten characterizations of linguistic theory in relation to those predictions, or naıve metaphorical proxies for linguistic theory, such as the retrieval of knowledge from a computer's memory, or assigning distributions to sentences as points in a discrete set of outcomes, rather than as points in a continuous, albeit inscrutable, latent semantic space.

A case in point is the resurgence of the notion of an "influence function" (Hampel, 1974), which attempts to assign weight to training sentences that an erroneous, indiscreet or salacious output can then be traced back to. Until very recently, the use of influence functions in LLMs was not computationally feasible. Now that it is somewhat feasible, the question is what it makes sense to do with them. In particular, the authors of these several papers on optimization and approximation of influence functions apparently never considered whether influence was merely a direct consequence of semantic similarity, a topic with a long history of proposed quantitative methods.

The central claim of this paper is that a better understanding of the potential of influence functions is attainable with a slightly less superficial understanding of linguistic theory. In particular, as a complement to the task of directly computing the semantic similarity of two expressions, we introduce the task of *entity invariance*, in which two related sentences are examined relative to a semantic argument that they share. The relation between these two sentences is composed of *grammatical transformations*, a now rather antiquated term for regular, meaning-preserving correspondences (at least in a reading that equates meaning with thematic role assignment) between syntactic forms. Passivization, topicalization and clefting are examples of transformations. (Chomsky, 1965) (Lambrecht, 2001) (Aelbrecht and Haegeman, 2012).

We describe a series of experiments and descriptive hypothesis tests which demonstrate that, under certain conditions, influence functions have a greater potential for invariance to syntactic transformations than conventional sentence embeddings in large-dimensional vector spaces. Just as in computer vision, where the ability to identify a shape is naturally tested for translation and rotation invariance, we assert that a semantic representation should be tested for invariance to diathesis and other syntactic transformations that ostensibly preserve meaning.

2 Methods

2.1 Sentence-BERT

As a canonical example of sentence embeddings, we select all-mpnet-base-v2 (Reimers and Gurevych, 2019), a sentence-transformer model that encodes sentences into a 768-dimensional dense vector space. The underlying model is the Microsoft mpnet-base model, pre-trained with the MPNet objective function (Song et al., 2020):

$$\mathbb{E}_{z \in Z} \sum_{t=c+1}^{n} \log P(x_{z_t} | x_{z_{\le c}}, M_{z_{>c}}; \theta) \tag{1}$$

This is a unified pre-training objective for both Masked Language Modeling (Devlin et al., 2019) and Permuted Language Modeling (Yang et al., 2019), inheriting the strengths of both. A sequence is permuted, and its right-most tokens are masked. The goal is then to predict the value of the masked token conditioned on all tokens preceding it, $x_{z_{\le c}}$, and the positions of the other masked tokens $M_{z_{>c}}$.

The model is then contrastively fine-tuned between sentence pairs in batches by computing the cosine similarities of their embeddings and comparing the cross-entropy loss with true pairs. The cosine similarities produce a value from -1 to 1. The cross-entropy loss then encourages the true pairs to have a larger value (closer to 1) while the non-pairs have a smaller one (closer to -1).

The resulting model accepts a sentence or paragraph and produces a vector encoding that captures some semantically relevant information.

2.2 Influence Functions

Influence functions are an older idea from statistics, re-introduced only recently to deep learning (Koh and Liang, 2017). Suppose there is a training dataset $D = \{z_i\}_{i=1}^{N}$ and a model with parameters $\theta \in \mathbb{R}^D$, fit using a loss function $\mathcal{L}$:

$$\theta^* = \underset{\theta \in \mathbb{R}^D}{\arg\min} \mathcal{J}(\theta, D) = \underset{\theta \in \mathbb{R}^D}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(z_i, \theta). \tag{2}$$

With this, we would like to investigate the effect of adding or removing a single training example z_m on the optimal parameters θ^*. By weighting that new training example by ϵ, we can describe the new optimal parameters with an additional training example as:

$$\theta^*(\epsilon) = \underset{\theta \in \mathbb{R}^D}{\arg\min} \mathcal{J}(\theta, D_\epsilon) \tag{3}$$

$$= \underset{\theta \in \mathbb{R}^D}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(z_i, \theta) + \epsilon \mathcal{L}(z_m, \theta). \tag{4}$$

Influence is defined as the first-order Taylor approximation to this function evaluated at $\epsilon = 0$. Using the Implicit Function Theorem, this is:

$$\mathcal{I}_{\theta^*}(z_m) = -H^{-1} \nabla_\theta \mathcal{L}(z_m, \theta^*) \tag{5}$$

where $H = \nabla_\theta^2 \mathcal{J}(\theta^*, D)$ is the Hessian of the empirical-loss function with the original dataset.

Since $\mathcal{I}_{\theta^*}$ is the linear approximation at 0, we can approximate the change in parameters as follows:

$$\theta^*(\epsilon) - \theta^* \approx \mathcal{I}_{\theta^*}(z_m)\epsilon \tag{6}$$

$$= -H^{-1} \nabla_\theta \mathcal{L}(z_m, \theta^*)\epsilon \tag{7}$$

Now, when we set $\epsilon = -\frac{1}{N}$ for some datapoint z_m already in the dataset, this corresponds to the effect of removing that datapoint.

Lastly, a change in parameters is difficult to interpret, so typically influence is measured on a more meaningful quantity such as validation loss or perplexity. Luckily, this can easily be done for any quantity $f(\theta)$ using the chain rule. For any meaningful measure f:

$$\mathcal{I}_f(z_m) = \nabla_\theta f(\theta^*)^T \mathcal{I}_{\theta^*}(z_m) \tag{8}$$

$$= -\nabla_\theta f(\theta^*)^T H^{-1} \nabla_\theta \mathcal{L}(z_m, \theta^*) \tag{9}$$

Applying $I_f(z_m)$ in the same way as before, we can approximate the change in this measure f due to the addition/removal of a datapoint with the following:

$$f(\theta^*(\epsilon)) - f(\theta^*) \approx \mathcal{I}_f(z_m)\epsilon \tag{10}$$

$$= -\nabla_\theta f(\theta^*)^T H^{-1} \nabla_\theta \mathcal{L}(z_m, \theta^*)\epsilon. \tag{11}$$

2.2.1 Influence in the Domain of LLMs

While influence functions are an old idea, numerous limitations kept them from being practical when examining neural-network-based architectures (Bae et al., 2024) (Zhang and Zhang, 2022) (Basu et al., 2021).

- Loss landscapes are not fully convex, meaning that the Hessian can be singular (thus it has no inverse).

- Even if the loss landscape were convex, the formulation of these objective functions implicitly assumes the model is trained to full convergence, which is almost never the case.
- Even if neither of these were an issue, the task of inverting the Hessian is by itself time-consuming.

These limitations have, over time, been addressed (Martens and Grosse, 2015) (George et al., 2018) (Martens, 2020) (Bae et al., 2024), mainly with clever approximations. The final result is then a reasonably efficient method for calculating influence for analyzing even large language models (Li et al., 2024), which we employ for our experiments. For a more detailed explanation, we refer the reader to (Grosse et al., 2023).

2.2.2 Influence for Language Modeling and Transformers

To use influence on the language-modelling task, we simply set the quantity f to be the following:

$$f(\theta) = \log p(z_c; \theta) \tag{12}$$

where z_c is the model's output and θ are the parameters of the transformer model. We follow previous work and use GPT2 (Radford et al., 2019) as the model to analyze, for which this log-likelihood decomposes using Bayes's Rule. Then the influence function approximates the instantaneous change in log-likelihood of generating an output z_c when removing or adding a piece of training data. For example, when a model generates, "Pythagoras was a ...", the presence of a training datapoint like "the Pythagorean theorem ..." is intuitively more important to this prediction than something less related like "The doctor suggested ...". Influence allows us to quantify the effect of a single datapoint from the training set by ablating it.

3 Problem Description

We investigate two capacities that we conjecture to be desirable of any model that aspires to true semantic reasoning: the now very well-studied ability to calculate the similarity in meaning between two sentences, and an invariance to meaning-preserving syntactic transformations.

In particular, we define *entity invariance* as a three-way comparison in which the congruence of the (now, usually a vector) representation of a fixed referring expression is calculated with a sentence that uses it, but relative to a baseline in which the same congruence is calculated between the same referring expression and a different but closely related sentence. For example, while the precise geometric relationship between the designator *John* and *John threw the ball* may be mostly inscrutable within modern neural vector representations of word and sentence meaning, we are perhaps justified in expecting that this relationship, whatever it is, will be the same as the one between *John* and *The ball was thrown by John*, *The ball, John threw* or *It is the ball that John threw,* because these various transformations are ostensibly meaning-preserving. This is a higher-order alternative to directly calculating the sentence similarity between the representations of *John threw the ball* and *The ball was thrown by John*, viewed through the lens of the meaning of *John*.

This has further implications with respect to phenomena like semantic masking (Shi and Penn, 2025), in which asymmetries have been observed in the ability of a document's context to obscure various inserted passages of text in question-answering tasks with LLMs. Rephrasing under a meaning-preserving transformation can actually alter these effects if the entity answer to a factoid question is not invariant to its sentence location.

The motivation behind both tasks is the same: given some semantics-related task, when replacing a sentence with a semantically equivalent yet syntactically transformed alternative, it should be the case that any method that claims to encode semantics should be robust to this replacement. Essentially, we claim moving across semantics-preserving transformations should not change the behavior of a true measure of semantics. For example, if sentence A is similar to sentence B according to some measure, and A' is the passivized form of A, then A' should be equally similar to sentence B. This is the sentence similarity task. If the subject of sentence A is deemed important by some measure, then the importance of that same subject on the sentence A' should also be equally important by that measure. This is the entity invariance task. We can approach both tasks with the aforementioned semantic tools: cosines of sBERT vectors and influence functions. An example is illustrated in Figure 1.

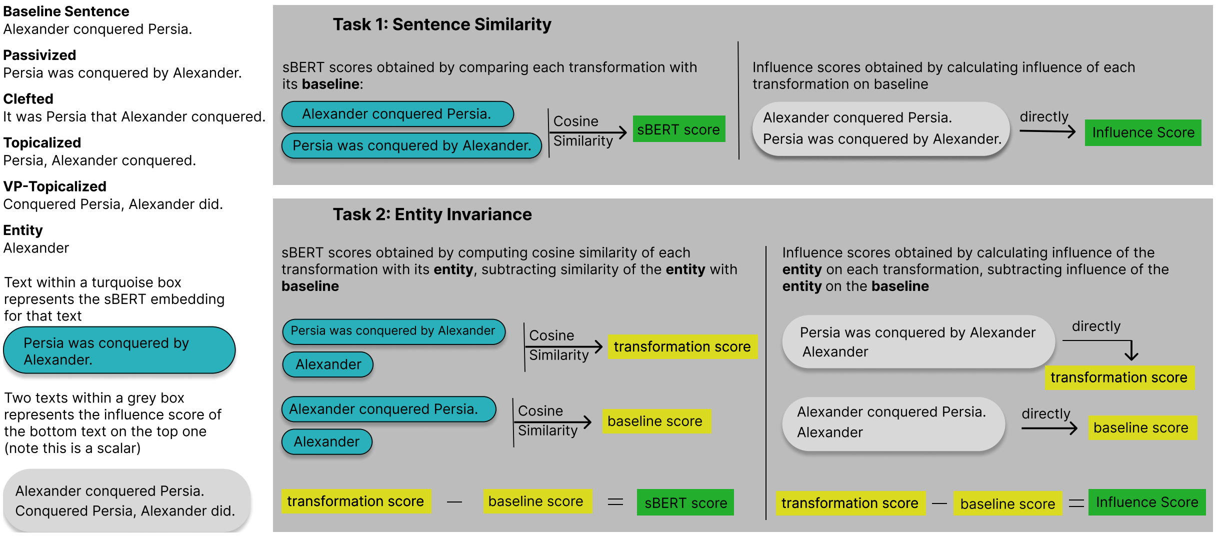
Baseline Sentence
Alexander conquered Persia.
Passivized
Persia was conquered by Alexander.
Clefted
It was Persia that Alexander conquered.
Topicalized
Persia, Alexander conquered.
VP-Topicalized
Conquered Persia, Alexander did.
Entity
Alexander
Text within a turquoise box represents the sBERT embedding for that text
Persia was conquered by Alexander.
Two texts within a grey box represents the influence score of the bottom text on the top one (note this is a scalar)
Alexander conquered Persia.
Conquered Persia, Alexander did.
Task 1: Sentence Similarity
sBERT scores obtained by comparing each transformation with its baseline:
Alexander conquered Persia.
Persia was conquered by Alexander.
Cosine Similarity
sBERT score
Influence scores obtained by calculating influence of each transformation on baseline
Alexander conquered Persia.
Persia was conquered by Alexander.
directly
Influence Score
Task 2: Entity Invariance
sBERT scores obtained by computing cosine similarity of each transformation with its entity, subtracting similarity of the entity with baseline
Persia was conquered by Alexander
Alexander
Cosine Similarity
transformation score
Alexander conquered Persia.
Alexander
Cosine Similarity
baseline score
transformation score − baseline score = sBERT score
Influence scores obtained by calculating influence of the entity on each transformation, subtracting influence of the entity on the baseline
Persia was conquered by Alexander
Alexander
directly
transformation score
Alexander conquered Persia.
Alexander
directly
baseline score
transformation score − baseline score = Influence Score

Figure 1: Example of both tasks under both metrics. Above shows the walkthrough of getting scores for the sentence "Alexander conquered Persia." in the passivized transformation. The above calculations are repeated for each transformation and for each sentence.

4 Experimental Setup

4.1 Datasets

4.1.1 Sentence Sampling and the Grammatical Transformation Dataset

In order to investigate grammar transformations on our semantic tasks in a controlled manner, we created a dataset that contains 50 random factual statements expressed in a simple sentence, containing only one independent clause. We prompt ChatGPT to produce a list of fact statements that are expressed in a simple sentence. We take 50 of these sentences as our baseline, with hand-filtering to remove any strange or duplicate sentences. We then prompt ChatGPT with these baselines and for each baseline, ask it to give a topicalizaed, clefted, vp-topicalized, and passivized form. Again, a final step involves meticulously going through the transformations to ensure accuracy. See Appendix C for details about the prompts. The result is a dataset containing 50 sentences in their base form. For each base form, a passivized, clefted, topicalized, and vp-topicalized form makes up the complete dataset. Refer to Table 1 for an example of an entry in the dataset, and refer to Appendix A for all the baseline sentences in the dataset (their transformations follow naturally).

Baseline	Alexander conquered Persia.
Passivization	Persia was conquered by Alexander.
Clefting	It was Persia that Alexander conquered.
Topicalization	Persia, Alexander conquered.
VP-Topicalization	Conquered Persia, Alexander did.

Table 1: One entry of the Grammatical Transformation Dataset

Baseline	Zorvik climbed Everest.
Passivization	Everest was climbed by Zorvik.
Clefting	It was Everest that Zorvik climbed.
Topicalization	Everest, Zorvik climbed.
VP-Topicalization	Climbed Everest, Zorvik did.

Table 2: One entry of the Made-Up Entity Dataset

4.1.2 Wikitext Dataset

The WikiText dataset (Merity et al., 2016) is a collection of over 100 million tokens taken from "good," i.e., featured articles in Wikipedia. Several earlier papers on influence functions have chosen to use this source, and so we have followed suit.

Influence functions are rather anomalous with respect to language modeling experiments. The language model (GPT2, in our case) is pre-trained on a large dataset D, but then it must also be fine-tuned on a smaller dataset with respect to the same language modelling objective. The influence calculations then determine how influential a certain training instance in the fine-tune dataset is on the generation of a query. This fine-tuned set exists only so that influence will not need to be computed on the entire pre-training dataset, which is massive.

We use the training partition of wikitext-2-raw-v1 as the basis of our fine-tuning set. Into this, we have inserted grammatically transformed sentences from the Grammatical Transformation dataset that are semantically unrelated to the wikitext that they are embedded in.

4.1.3 Made-Up Entity Dataset

But because the pre-trained model may have seen some version of the same data, it does make sense to have another dataset where we rename all entities that appear as subjects in the corresponding, untransformed baseline sentences (the transformations then typically change which grammatical function that entity will have) with completely made-up entities. When we use these renamed, baseline sentences as queries during influence calculations, we can then reasonably be assured that the influence will have come mainly from the correspondingly renamed transformation in the fine-tuned set.

Table 2 shows an entry in the Made-Up Entity dataset, and Appendix A shows all the made-up entities.

4.2 Calculating Sentence Similarity

In directly calculating sentence similarity with sBERT vectors, we simply compute the cosine of the sBERT encoding of a baseline sentence with that of each of its transformations in the Grammatical Transformation Dataset in turn. We used the sentence transformer all-mpnet-base-v2 described in Section 2.1.

When calculating sentence similarity with influence functions, we assume that sentences that are more similar will be more influential. Our made-up entity dataset has been concocted with nonsense names so that the transformed sentence that was inserted into the fine-tuning text will, in spite of its transformation, be the most semantically similar. The influence score will then correspond to how similar they are.

Note that due to the symmetry built into the definition of influence functions, we do not need to explicitly symmetrically close our definition of similarity here.

To support batched calculations, all of our added entries are padded to 20 tokens, long enough to cover the longest transformed sentence in our dataset. With this setup, we can obtain the influence of each transformation on generating its own baseline variant.

4.3 Calculating Entity Invariance

When using sBERT to calculate entity invariance, we calculate:

$$\frac{(e \cdot t)}{|e||t|} - \frac{(e \cdot b)}{|e||b|}$$

where e, t and b are the sBERT vectors for the entity, transformed sentence and untransformed baseline, respectively. Note that this calculation avails itself of sBERT's indifference to the semantic type of its input.

We do this for each transformed sentence, for each entry in the Grammatical Transformation dataset.

With influence functions, we again assume that the congruence or salience of an entity to a particular text will be reflected by a greater influence. We again avail ourselves of influence's indifference to the semantic type of the query, which can be as simple as a referring expression. In our experiments, the entity in question will always be the subject of the untransformed baseline sentence. We subtract the influence of the entity on the baseline from the influence of the entity on the transformed sentence. Padding is the same as with sentence similarity. Figure 1 presents an example of both tasks under both metrics.

5 Results and Findings

Let us first begin by noting that, across both tasks and all syntactic transformations, there is a tight, linear correspondence between sBERT vector cosines and influence scores. Their Pearson correlation is 0.9326, with a p-value of $2.62 \times 10.^{-178}$

As for the specific grammatical transformations, the five rows shown in the tables in this section were chosen because they represent overall trends; the full results for all 50 sentences can be found in Appendix B. In addition, influence scores were scaled with arctan, compressing the range to $-\pi/2$ to $\pi/2$.

Table 3 contains sentence similarity scores using sBERT cosines. For the sentence similarity task, sBERT tends to encode the passivized forms of sentences most similarly to their corresponding baseline sentences. Table 4 contains sentence similarity scores using influence functions. In stark contrast to the sBERT results, influence finds both topicalizations to be most similar to their baselines, whereas passivization is the least similar. In both tables, we can see that the scores are near the top of their respective scales.

Passivization	Clefting	Topicalization	VP-Topicalization
0.9325544834	0.8652806878	0.8628834486	0.90064466
0.9408032894	0.9085036516	0.883110702	0.8844070435
0.9199316502	0.8648024201	0.8657934666	0.8941929936
0.9520395398	0.9430727363	0.9146342278	0.9339743257
0.9660890102	0.8314833641	0.8642077446	0.9086657166

Table 3: Scores of the Sentence Similarity task between the baseline and each of the different transformations using sBERT cosine similarities. Each row corresponds to one row in the Grammatical Transformation dataset, and each column to a grammatical transformation. Note for this and all tables using this color pattern, white represents the smallest value, and dark green is the largest. Rows are independently colour mapped.

Passivization	Clefting	Topicalization	VP-topicalization
1.570795547	1.570795942	1.570796084	1.570796024
1.570796187	1.570796237	1.570796266	1.570796248
1.57079604	1.570796097	1.57079621	1.570796157
1.570795623	1.570796153	1.570796207	1.570796074
1.570793101	1.570796037	1.570796129	1.570796194

Table 4: Scores of the Sentence Similarity task between the baseline and each of the different transformations using influence functions. The scores have been normalized using arctangents.

Table 5 shows the entity invariance scores using sBERT cosines. For this task, sBERT vectors are most invariant to passivization relative to their encodings of the respective baseline sentence, whereas clefting exhibits the most variance. Table 6 shows the entity invariance scores using influence functions. For this particular combination, it is more difficult to spot any sort of trend or preference for one transformation over the others. Both of these scores are difference calculations. In the case of sBERT, the differences are closely range-bound around zero, meaning that the effect of using any grammatical transformation was minimal. In the case of influence functions, the prominence of values near $-\pi/2$ shows that all of the grammatical transformations we experimented with resulted in a suppression of influence scores relative to the baseline subject.

Passivization	Clefting	Topicalization	VP-topicalization
-0.09655714035	-0.1692547202	-0.04888242483	-0.08329671621
0.03376698494	-0.01508197188	0.02062654495	-0.01081442833
-0.09354573488	-0.1332816482	-0.1059363484	-0.1363123655
-0.02574926615	-0.05807337165	-0.0239841342	-0.07041674852
-0.02594101429	-0.09438753128	-0.04366868734	-0.07553547621

Table 5: Scores of the Entity Invariance task between the subject of the baseline sentence and each of that sentence's different transformations using sBERT cosines.

Passivization	Clefting	Topicalization	VP-topicalization
-1.570795954	-1.570796167	-1.570795853	-1.570795884
-1.570796289	-1.570796284	-1.570796288	-1.570796289
-1.570796251	-1.570796245	-1.570796264	-1.570796272
-1.570796283	-1.570796278	-1.570796279	-1.570796278
-1.570796223	-1.570796245	-1.570796238	-1.57079623

Table 6: Scores of the Entity Invariance task between the baseline subject relative to each transformation using influence functions. The scores have been normalized using arctangents.

5.1 Significance of Grammatical Transformations

It is also possible to examine differences in the effects of the four grammatical transformations that we selected. The distributions of the various scores across tasks, both jointly and severally, fail Levene's test of homoscedasticity, so a repeated-measures Friedman's test is the appropriate way to test for significant differences among their medians. Its null hypothesis is that there is no significant difference among the four transformations, which would imply (but not prove) a degree of resilience in the chosen semantic representation. As shown in Table 7, the choice of grammatical transformation is significant in the direct sentence similarity task, regardless of method, but is significant for entity invariance only with sBERT cosines, not with influence functions. Note that the magnitudes of the p-values are at opposite poles, so this is a matter of kind, not degree. Table 8 shows the respective test statistics with their effect sizes. The three significant effect sizes are all considered large, because they are greater than 0.5.

For the settings found to be statistically significant, we present a ranking of transformation preference (higher scores are more preferred) in Table 9. This confirms that for the task of sentence similarity, influence finds passivization to produce the least similarity, and therefore the most difference in meaning, while sBERT finds passivization to be most similar. In fact, while they have similar p-values and test statistics to those for sBERT vector cosines, their ranking of grammatical transformations is the exact opposite.

For entity invariance, on the other hand, sBERT once again finds passivization to best preserve it, although clefting preserves it the least. In both tasks, we are of course not testing whether passivization influences meaning, but rather, given that passivization is thought to be meaning-preserving, whether sBERT cosines and influence functions perform as we want them to.

	Sentence Similarity	Entity Invariance
sBERT	2.32×10^{-15}	3.97×10^{-7}
Influence	1.69×10^{-15}	0.983

Table 7: p-values of Friedman's test for different experimental settings.

	Sentence Similarity	Entity Invariance
sBERT	71.23 / 1.1936	32.57 / 0.807
Influence	71.88 / 1.199	0.17 / *0.058

Table 8: Test statistics (χ^2) / effect sizes (ϕ) of Friedman's test for different experimental settings (the lower-right effect size is hypothetical, as no significance has been demonstrated).

5.2 Effect of Concocted Names

As shown in Table 10, the effect of concocting the names of the fixed entities magnifies the effect of changing the grammatical transformation in the entity invariance task to the point that it becomes statistically significant, and of moderate, almost large size.

6 Discussion

That influence functions might demonstrate any resilience to syntactic transformations is indeed interesting, because: (1) sBERT vectors do not (nor does any other vector-based representational scheme that we are aware of), in spite of how amazingly well they work as semantic representations, and (2) it means that influence functions bring us that much closer to being able to truly work with the meanings of sentences rather than more superficial aspects of their syntactic realizations. Nevertheless, this resilience has only been seen in our examination of something more subtle, where we look not at differences in meaning, but differences in influence scores relative to a fixed entity, and thus arguably differences in differences in meaning. Were it not for entity invariance, in fact, one might wonder why influence scores bothered to exist, given their strong Pearson correlations to sBERT-vector cosines and fickleness with respect to syntactic transformations in more direct comparisons of sentence meaning.

The results on the Made-Up Entity dataset suggest that at least some of the resilience of influence functions is due to their ability to draw upon the meanings of the pre-trained data or the syntactic variety of their expression, or both, in order to see through the effects of a syntactic transformation. In typical LLM fashion, however, the patterns learned by the language model in relation to this are not sufficiently robust or principled to withstand, for example, an innocuous change in the semantic arguments. And so an innovation that was designed to isolate the effects of the query around the transformed sentence in fact hurt performance.

6.1 Limitations

We cannot flatly claim that influence functions are a better alternative to sBERT vectors, in part because of the adverse effects of consistently changing names. There are other limitations, too, the chief of which is that sBERT uses an encoder-style model which contains bi-directional context, while the Anthropic code base and paper for influence functions is focused around GPT2 and other decoder models that only see previous tokens in its history. So it is impossible to determine the extent to which the entity invariance we saw with influence functions is due to the underlying decoder architecture without rewriting that code base. What we can already affirm is that this difference in architecture was not enough for influence scores to fall out of lock step with sBERT cosines in the Pearson correlation test that we conducted.

Another limitation is our choice of a small number of grammatical transformations for experimentation. The results presented in Table 9 naturally single out passivization from the other transformations, and indeed passivization is special. It is the only transformation among the four that we selected to unequivocally constitute A-movement, and the only one that rotates the grammatical function assignment around the arguments of the baseline sentence. It is also the only one of the four that has overt morphological reflexes, although both clefting and VP-clefting would also cause the LLM's tokenizer to change the length of the input. One might also argue that certain of these four transformations are easier to withstand or easier to predict the consequences of, using the measurement tools at our disposal, either because of the structural complexity of the transformation in terms of a chosen syntactic representation, or because of a variance in their relative frequencies in the pretraining corpus. We would, at the same time, like to expand the experimental list of transformations, while better balancing these other effects, but these two purposes work against each other.

	Passivization	Clefting	Topicalization	VP-Topic
Influence on Sentence Similarity	1.570795571	1.570795892	1.570796019	1.570796057
sBERT on Sentence Similarity	0.937302351	0.9078437984	0.8857396245	0.8987811208
sBERT on Entity Invariance	-0.05444133282	-0.08711430431	-0.05307358504	-0.07616019249

Table 9: Medians of the scores on the Grammatical Transformation dataset for each transformation under statistically significant conditions, ranked by colour.

	p-values	Test Statistics	Effect Size
Sentence Similarity	3.79×10^{-14}	65.568	1.145
Entity Invariance	0.008	11.712	0.484

Table 10: p-values, test statistics (χ^2), and effect sizes (ϕ) for different tasks with the Made-Up Entity dataset (influence functions only).

7 Conclusion

Along with neural language models has come increasing concern over transparency and explainability. Influence functions are one example of an attempt to understand or interpret language models. There is some evidence, as shown in this paper, that influence functions are good for more than assigning blame for faulty output. They correlate well with sentence-similarity scores.

Using entity invariance over grammatical transformations, we have been able to distinguish the two, however. While sentence embeddings are not resilient to syntactic transformations in any of our experimental settings, in certain conditions, influence functions are. This is important, because meaning representations should be invariant to meaning-preserving transformations.

It will be important to repeat this experiment after reworking either the Anthropic code base or sBERT so that they can run on the same kind of model. It will also be important to expand and better control the list of syntactic transformations.

References

Lobke Aelbrecht and Liliane Haegeman. 2012. Vp-ellipsis is not licensed by vp-topicalization. *Linguistic Inquiry*, 43(4):591–614.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger Grosse. 2024. If influence functions are the answer, then what is the question? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Samyadeep Basu, Phillip Pope, and Soheil Feizi. 2021. Influence functions in deep learning are fragile. In *ICLR*. OpenReview.net.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. 2018. Fast approximate natural gradient descent in a kronecker-factored eigenbasis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9573–9583, Red Hook, NY, USA. Curran Associates Inc.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. FAR.ai Alignment Workshop 2023.

Frank R. Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1885–1894. JMLR.org.

Knud Lambrecht. 2001. A framework for the analysis of cleft constructions. *Linguistics*, 39(3):463.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. 2024. Do influence functions work on large language models? *Preprint*, arXiv:2409.19998.

James Martens. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76.

James Martens and Roger Grosse. 2015. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2408–2417. JMLR.org.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ken Shi and Gerald Penn. 2025. Semantic masking in a needle-in-a-haystack test for evaluating large language model long-text capabilities. In *Proceedings of the Writing Aids at the Crossroads of AI, Cognitive Science and NLP WR-AI-CogS Workshop at the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.

Rui Zhang and Shihua Zhang. 2022. Rethinking influence functions of neural networks in the over-parameterized regime. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):9082–9090.

A Full Datasets

Table 11 contains all the concocted entities that, after replacing the subjects in the Grammatical Transformation dataset, form the Made-Up Entity dataset described in 4.1.3.

Table 12 contains all the baseline sentences in the Grammatical Transformation dataset described in Section 4.1.1. The transformations are omitted for brevity but follow directly from the baselines.

B Full Result Tables

Table 13 contains the full results of the sentence similarity task on both metrics. Note that there are 50 rows of data, each corresponding to an entry in the Grammatical Transformation Dataset. Colors are mapped such that the smallest is white, largest is dark green and intermediate values are gradated uniformly, and in addition, each row is done independently.

Table 14 contains the full results of the entity invariance task on both metrics.

Table 15 contains the full results for the test of influence on the Made-Up Entity dataset on both tasks, detailed in Section 4.1.3.

C Prompts

Prompt to generate the baseline fact sentences: *Please provide me a list of factual statements like "Mozart composed symphonies" that follow the simple sentence structure.*

Given the list of baseline sentences, the prompt to generate a transformation: *I will provide you a list of sentences. You are to take each sentence and topicalize it. For example, if given "John liked Mary." you are to return "Mary, John liked."*. The same prompt can be adapted for the other transformations.

Made-up Entities in the Made-Up Entity Dataset

Kolpytimia	Fervan	Phran	Zorvik	Ivoren
Jymilopy	Galros	Quirin	Reilktyia	Jexar
Fulkingra	Hivian	Raxen	Avaron	Kynor
Liuntmat	Ivren	Salven	Brenix	Larven
Kolparop	Jovik	Torvin	Cyrin	Morden
Funmilip	Kelrin	Uvorn	Dralin	Nexor
Belrix	Laxor	Vexan	Elvir	Shadrin
Cevran	Merin	Wavric	Fixon	Fullinma
Darvon	Novin	Xalden	Gravin	Dilkop
Emlian	Orvex	Yavren	Haldor	Imnity

Table 11: All entities in our Made-Up Entity Dataset. These replace the subjects of the Grammatical Transformation dataset to form the new dataset

Baseline Sentences in the Grammatical Transformation Dataset

Albert Einstein developed the theory of relativity.	Armstrong landed on the moon.
Isaac Newton formulated the laws of motion.	Fleming discovered penicillin.
Leonardo da Vinci painted the Mona Lisa.	Darwin explained evolution.
William Shakespeare wrote Hamlet.	Jobs founded Apple.
Marie Curie discovered radium.	Beethoven composed Fur Elise.
J.K. Rowling wrote the Harry Potter series.	Hillary climbed Everest.
Vincent van Gogh painted The Starry Night.	Pasteur developed vaccines.
Nikola Tesla invented the alternating current (AC) motor.	Galileo built telescopes.
Georgy Zhukov led the defense of Stalingrad.	Ford revolutionized manufacturing.
Alexander Fleming discovered penicillin.	Orwell wrote 1984.
Michelangelo sculpted David.	Picasso painted Guernica.
Charles Darwin developed the theory of evolution.	Edison patented the light bulb.
Thomas Edison invented the electric light bulb.	Mandela fought apartheid.
Beethoven composed Symphony No. 5.	Turing cracked the Enigma code.
Alexander Graham Bell invented the telephone.	Pythagoras discovered the Pythagorean theorem.
Mozart composed The Magic Flute.	Hitchcock directed Psycho.
Leonardo DiCaprio played the role of Jay Gatsby.	Mozart composed Don Giovanni.
Columbus discovered America.	Washington led the Continental Army.
The Wright brothers invented the airplane.	Napoleon invaded Russia.
Alexander conquered Persia.	Franklin invented the lightning rod.
Marie Curie studied radioactivity.	Curie discovered polonium.
Tesla designed alternating current systems.	Kepler described planetary motion.
The Romans built aqueducts.	Gagarin orbited Earth.
Magellan circumnavigated the globe.	Caesar crossed the Rubicon.
Gutenberg invented the printing press.	Chopin composed nocturnes.

Table 12: All 50 baseline sentences used in the Grammatical Transformation Dataset. Not included for brevity are the corresponding grammatical transformations but they all follow naturally to make up the full dataset.

Sentence Similarity via Influence				Sentence Similarity via Sentence-BERT			
Passivization	**Clefting**	**Topicalization**	**VP-Topicalization**	**Passivization**	**Clefting**	**Topicalization**	**VP-Topicalization**
1281673.25	2596264	4117991.25	3298893.25	0.9325544834	0.8652806878	0.8628834486	0.90064466
5032188.5	7685314	6797391	6382516	0.9408032894	0.9085036516	0.883110702	0.8844070435
7136764.5	11084613	16475059	12699568	0.9306652546	0.8681627512	0.869109869	0.8732848167
3492748	4352391	8556159	5876025	0.9199316502	0.8648024201	0.8657934666	0.8941929936
1532958	2286936.75	1962189.875	2141318.75	0.9520395398	0.9430727363	0.9146342278	0.9339743257
4620397	13047812	8191310.5	20385430	0.9660890102	0.8314833641	0.8642077446	0.9086657166
1275383.625	2828912.75	4700864.5	6202506.5	0.9333539009	0.9185542464	0.9130138755	0.8701131344
3085148.25	2570608.75	3505769.5	5189386.5	0.9549874067	0.9168089628	0.9037501812	0.9442888498
5773955.5	7070393.5	10300728	12519818	0.9459875226	0.9487894773	0.9275122881	0.8671823144
4353710.5	3066081.75	5515658	3065397.75	0.9545772076	0.9427666068	0.9304442406	0.9420560598
1201822.25	762331.3125	6281364	5113827	0.9067315459	0.9071839452	0.8697237372	0.902159512
6461203	5040498.5	7943203.5	8796788	0.9197968245	0.8237189054	0.8344243169	0.8498998284
1900646.625	1747893.375	3532316.25	5013893	0.9284735918	0.84999156	0.8362667561	0.9234173894
3811932	5775193	5407366.5	8614384	0.9404629469	0.940613687	0.9342517853	0.8057485819
8396329	13875857	38902156	-336879.2813	0.9341417551	0.8465870023	0.7667613029	0.8968443274
2837912.5	2746652.75	3477327.5	5668458	0.9407648444	0.7814874649	0.8770526648	0.8991389275
2148719.75	2310032.5	3563150	7584823.5	0.9522520304	0.9452135563	0.8985278606	0.836519599
1163579.875	1002357.438	2497049.5	1499065.875	0.9046645164	0.8813423514	0.7317293882	0.8887551427
1367728.375	614765.3125	2399663.5	2348435.25	0.9161099195	0.8231762052	0.8400527835	0.8983151913
1972371.625	5103495.5	3946486.5	6394397.5	0.9530593753	0.918993175	0.8561660051	0.9100579023
712751.5	1288459.25	2425012.75	3654488	0.9562042356	0.9505699277	0.9037286043	0.8873476982
3226312.25	12242624	3931645.75	13437938	0.9615622759	0.9444450736	0.9279776812	0.8876610994
1420691.625	5752419	8330069	3952169	0.9537521601	0.9556134939	0.9316477776	0.8822870851
309985.2188	3447946	5053616	7553659	0.9246538281	0.8528832197	0.856222868	0.9120983481
800681	538536.625	2079399.75	2897636.25	0.9088691473	0.8832570314	0.8298295736	0.8844642043
91632.98438	1477246.75	1883422.875	2362679.5	0.8620303273	0.7549761534	0.7469062209	0.8201477528
2736904.5	441928.3438	763860.5	1111945.375	0.9550385475	0.9451751709	0.9499857426	0.9289374352
317121.375	905179.6875	2216923.25	1574170.75	0.8973581791	0.8525787592	0.8231647611	0.8734014034
1058726.875	2258758.75	3243198.25	4167219.75	0.9199647903	0.8848507404	0.8137908578	0.903968513
-1187882.375	6206952.5	1382713.875	2774814.75	0.9419152141	0.9371224046	0.8946403861	0.9033447504
1384578.5	11605822	1470714.5	12923023	0.8947380781	0.8980829716	0.8914081454	0.878882587
398852.7188	877115.1875	2340028.25	1418243.75	0.9431471229	0.9407480955	0.927508533	0.8732652068
512260.9688	124985.6953	1507217.375	3009961.5	0.9419971704	0.9371962547	0.9451744556	0.9277190566
927574	-569455.8125	890655.3125	3681423	0.9550658464	0.9444385767	0.9141231775	0.9183707237
733503.875	1276804.125	3643654	613235	0.9289010763	0.8879346251	0.8347960711	0.8943598866
2617136.25	1795073.5	2855236	7847653	0.9220842123	0.915694356	0.8794906735	0.8984233141
636359.25	126831.7031	1154313.75	4479112.5	0.9173202515	0.9273391962	0.895643115	0.9292954206
688857.4375	870001.75	1613224.25	2188606.25	0.8809921145	0.8822927475	0.8651847839	0.8980981708
897886.75	2438285.5	3755775.25	5557944	0.9443784356	0.8753024936	0.8788477182	0.900886178
941638.5625	1710459.5	3206986.75	2542614.5	0.9473628402	0.9563817978	0.9277408123	0.9360141158
30088.2207	3036802.75	1362985.5	3273471.25	0.8767876625	0.8422478437	0.918872118	0.8392100334
-774782.5625	4324408	3250920.75	7688956.5	0.9330461621	0.922550559	0.888368547	0.9031774402
1543971.625	1883542.375	2346876.5	2069620.625	0.9551422596	0.9336919785	0.9073114991	0.9046003222
2149576.5	1670727.625	4511604.5	2736040.75	0.9525103569	0.8796239495	0.8039262891	0.8643612266
-66643.16406	208990.6719	760295.8125	2674516.25	0.9314661622	0.906768024	0.925755322	0.924367547
1650114.375	3372535	1811241.375	3720192	0.9604322314	0.9507023096	0.9522266388	0.9335971475
2179720.75	94931.34375	1805788.5	2851061.25	0.9632445574	0.9436131716	0.9506351948	0.9086754918
908247.5	2219431.75	6535197.5	2098055.5	0.9507032633	0.9038532972	0.8263111115	0.9095230699
-470488.875	688643.0625	2670438.5	3444965.75	0.9334220886	0.9068481922	0.8767338395	0.9065231085
-1013974.75	6141339	2260686	7763943	0.9230386019	0.9325930476	0.9073643088	0.8753144145

Table 13: Full results of Sentence Similarity for both metrics on the entire Grammatical Transformation Dataset

Entity invariance via Influence				Entity invariance via Sentence-BERT			
Passivization	**Clefting**	**Topicalization**	**VP-topicalization**	**Passivization**	**Clefting**	**Topicalization**	**VP-topicalization**
-2685623	-6254405	-2108614	-2258712	-0.07981181145	-0.1866739392	-0.1492462158	-0.07941681147
-26642611	-23641762	-25821433	-26574901	-0.06372123957	-0.1190310717	-0.112989068	-0.0492888093
-13220497	-12153932	-15876012	-18372530	-0.08093649149	-0.1657860875	-0.06755018234	-0.04545301199
-23089960.75	-20474648.5	-20957768	-20638588.5	-0.1250346899	-0.1121886373	-0.05282765627	-0.1104551554
-9629271.5	-12170401.5	-11296740.5	-10278239.5	-0.004707098007	-0.06935936213	-0.05996596813	0.01575297117
-27716518	-24282692	-25155550	-24522918	-0.02630108595	-0.1617150307	-0.1105882525	-0.1176011562
-25609095	-25669994	-24993926	-25406562	-0.0555267334	-0.07726699114	-0.0490463376	-0.05264532566
-14933946	-12275290	-16347746	-16882840	-0.08301895857	-0.100716114	-0.07386052608	-0.08472329378
-15794400	-8453728	-11933606	-11977778	0.003785192966	-0.02281010151	-0.06004858017	-0.1017688513
-30020002	-19234332	-26009050	-21754662	-0.04582571983	-0.06943738461	-0.06015014648	-0.02689957619
-48643260	-52031866	-48561260	-46328487	-0.1114506721	-0.09153693914	-0.05393457413	-0.1327273846
-13833647	-9833368	-13587115	-10178348	-0.06523412466	-0.1308091283	-0.08723050356	-0.06536006927
-26825080	-25047562	-26189696	-25613158	-0.09480243921	-0.1450120807	-0.1342134476	-0.05758196115
-1135976.25	-475462.5	-2344986.125	-1476259.547	-0.0925809741	-0.1185005307	-0.05976593494	-0.06406724453
-39184466	-38029180	-31495636	-36814442	-0.04791623354	-0.1590764523	-0.1918034554	-0.08462017775
-1247697.25	-1582920.375	-2108257.125	-1594213.875	-0.09655714035	-0.1692547202	-0.04888242483	-0.08329671621
-5411460	-6072064	-5347273	-5201701	0.03376698494	-0.01508197188	0.02062654495	-0.01081442833
-4791516.125	-3093396.25	-3368433.25	-3402382.5	-0.09742739797	-0.08908066154	-0.006914794445	-0.09745392203
-12588536	-28372232	-34465044	-39599843	-0.09354573488	-0.1332816482	-0.1059363484	-0.1363123655
-5319618.375	-4277966.313	-4288703.438	-5360857.469	-0.02574926615	-0.05807337165	-0.0239841342	-0.07041674852
-7629488	-11472516.5	-10119478.5	-9223493	-0.01981073618	-0.03779411316	-0.05990833044	-0.03124922514
-1823057.594	-1784169.672	-1747495.781	-1909726.992	-0.03912311792	-0.07392579317	0.03000319004	-0.06897968054
-31943223	-28038618	-34662588	-32163390	-0.05335593224	-0.03372785449	-0.01171341538	-0.02090236545
-1104358.875	-2584393.844	-3484716.25	-1595341.125	-0.02594101429	-0.09438753128	-0.04366868734	-0.07553547621
-3454303.375	-1707085.625	-1072567.125	-1806547.125	-0.06934568286	-0.07635483146	-0.07115519047	-0.1302825809
2332928.625	2165193.875	2180960.313	2484504.125	-0.08078327775	-0.1065143049	-0.1084765792	-0.1459647715
-12339632.25	-13939776.63	-16306948.5	-15179150.88	-0.03378689289	-0.02639275789	0.001132577658	0.02219408751
-2894744.125	-3501600.211	-3639278.414	-3362675.984	-0.04256004095	-0.02782595158	0.06826972961	-0.003503620625
-198664.1445	-228390.333	-185710.7344	-199948.0586	-0.06719768047	-0.1265891194	-0.01654732227	-0.07524868846
-414397.375	-730582.7031	-678280.2813	-781506.2969	-0.09697979689	-0.08204746246	-0.06582641602	-0.05791759491
-283525.1289	-284416.8887	-261344.5234	-260542.9336	-0.09624645114	-0.08741539717	-0.04021796584	-0.07799932361
-8322824	-7754106	-6927294	-7234931	-0.03954720497	-0.04791337252	-0.01593309641	-0.1249685585
-38572803	-24507834	-34113582	-31726072	-0.03267228603	-0.04540675879	-0.02843618393	-0.04879248142
-18167.93164	-18284.80469	-22466.16797	-15712.23633	-0.02734774351	-0.06580168009	-0.05387979746	-0.1557758152
-446892.125	-483367.3125	-479492.9375	-1020807	-0.07068240643	-0.1144337654	-0.09733355045	-0.07694244385
-3250265.25	-4003595.563	-3033394	-3140831.75	-0.1278484464	-0.1360321045	-0.0533195138	-0.09156519175
-1908188.188	-831025.9375	-792524	-1234735.563	-0.1284969449	-0.1011826992	-0.09552234411	-0.04509288073
-1874450.125	-1568216	-1688667.938	-1822333.188	-0.1489322186	-0.1448811293	-0.06250846386	-0.1114014387
-4128723.789	-2868800.25	-5321833.625	-1356737.25	-0.04154163599	-0.07667589188	-0.0476590395	-0.07678490877
-29504152	-32576832	-30273304	-28288826	-0.03211379051	-0.04234272242	0.02607136965	-0.01085174084
26074.9375	-451694.0859	-220842.4688	-318559.5	-0.03710752726	-0.1560547352	-0.09752297401	-0.08025348186
-1946600.25	-2015420.75	-2433695.125	-1350453.25	-0.0997890234	-0.02228420973	0.01836383343	-0.08943325281
-718451.7813	-944356.6641	-945010.3594	-978230.1328	0.02094578743	-0.02048495412	0.01983216405	-0.1224358678
-29409283.88	-30200487.56	-29657554.56	-29677078.81	0.01468878984	-0.08090877533	-0.007811784744	-0.05628025532
-1109870.063	-1860999.219	-1730642.281	-2111163.311	-0.04685598612	-0.05257755518	-0.0422347784	-0.04875138402
-2075367	-2253507.375	-2368550.75	-2355812.188	-0.04864227772	-0.0438978672	-0.04264587164	-0.02939426899
-326502	-2257342.25	-2389740.875	-1023074	-0.06806963682	-0.08834481239	-0.0005748867989	-0.1109085083
-17051838.84	-16849220.94	-15265631	-16327568.88	-0.03762674332	-0.08681321144	-0.09053331614	-0.1175132394
-2649977.906	-2593891.906	-2393907	-2046743.75	-0.02224761248	-0.04226249456	-0.0484764576	-0.06495755911
-26235076.38	-26820000.16	-26304973.25	-26184891	-0.097905159	-0.08832764626	-0.02254664898	-0.1072673202

Table 14: Full results of Entity invariance for both metrics on the entire Grammatical Transformation Dataset

Sentence Similarity via Influence				Entity Invariance via Influence			
Passivization	**Clefting**	**Topicalization**	**VP-topicalization**	**Passivization**	**Clefting**	**Topicalization**	**VP-topicalization**
20699412	18898032	18216750	14871124	-27748200	-21957264	-19352244	-16531028
16107982	16881070	22166466	18357252	-25357752	-31761387	-25464259	-24658304
8983601	9679158	13356021	15754497	-15623932.5	-15539081	-10364748	-19133420
6311177	31551756	15346468	33377176	-32253152	-34866796	-30644562	-31838056
-4400801	3247743.25	8152989	-1417421.375	-37055640	-34900592	-44443604	-41120728
7341449.5	4947732.5	10200429	8779436	-10611611	-9151762	-10131612	-9676926
4204987.5	3982068	3914922.75	14536122	-2995158.75	-2896915.563	-2880270.25	-3905569.219
39287480	39517636	41943084	43717972	-23887976.25	-24590565.5	-24768332.5	-24817787.5
5499680.5	5292868	9293386	14166252	-6988639.25	-8419173.469	-8611006.188	-8578112
795575.6875	721371.75	2425898	1754051.875	-16320768	-17970983	-18785290	-17531584
1111503.125	2716567.25	1668867.125	5971097	-46875189.25	-48718234.09	-48014607.44	-47222828.75
3494235.5	5168803	5021732.5	6507909	-1306780	604563.5	-1818692	-198710
2854848.5	2777719.5	3898743.25	5461473	-44338464	-56349592	-50853727	-52246354
1941477.5	9786865	5676007.5	9990937	2108757.125	2131060.875	1837815.625	1714763
806739.75	1078171	3864495.25	1805136.25	-3313877	-5654521	-6208624	-6319545
1971969.5	2315744.5	3322627	4030830	609971	-116435	82594.09375	-293907.0938
3113887.25	3309663.25	5848674	7088668	-61137592.38	-62781103.19	-61198772.5	-61969797.77
904815.0625	4273071.5	1094813.375	3007366	-4451510.205	-4695470.813	-4905691.969	-4977761.5
476680.4375	180478.6094	677784.3125	1363827.125	-8516562.813	-8411518.031	-8619159.438	-8473698.125
-1440849.125	6980230.5	5300396	4356878.5	-11607418	-14793079	-14087506.5	-15849524.5
1288198.375	3206311.25	5574628	6867961	3482046.375	3396289.5	2969881.625	2794664.625
3750752	9526400	4090863.75	12548599	1851081.109	1852303.016	1861345.625	1741095.734
1479134.5	2008157.5	1753084.5	1594822.125	-2386763.75	-2982937.656	-2696552.434	-2792250.531
622985.6875	2346483.75	6988319	7975886	-18440996.5	-23206848	-19082092.5	-22463687
537622.6875	960032	2324116.75	3750187.75	-1748337	-3400219.875	-2762614.938	-3739159.875
1378174.875	1122773.25	2615694	3123351.5	-34615689.75	-35094075.25	-34674898.25	-35754053.25
347643.125	594932.75	1796586.375	2670808.75	-11495633.88	-11424431.88	-11695386.09	-11622510.38
305673.8125	661324.9375	2004349.625	1520810.75	-4411322.625	-3197441.75	-4706700.25	-4590533
2249547	3402734.5	6386406	4276829.5	-25215807	-29706655	-25352677	-28966407
3405452.25	12879227	5271509	8707653	-11760492	-14378362	-13283870	-10874287
768798.9375	10446396	5316533.5	12891541	-20133013.5	-20565648.5	-21723378	-20852058
9640745	9694372	6758299.5	3705323.75	-16910541	-22972139	-16782637	-18804143
355867.75	1285882.375	1297126.75	1909627.625	344915.875	161834.25	420683.5	-502630.875
1240693.875	2390800	3266505.75	3750562.5	-84545504	-86513072	-95985880	-85100844
126504.9453	689580.875	1197900.25	1589813.875	-2965967.5	-2691341.563	-2771703.875	-2990364.906
10040059	7601738	3605134.75	6452776.5	-12713596	-14162309	-9163776	-13836231.5
574604.875	1275667.75	2521754	3318908.25	-241800.1602	-507822.1719	-1196177.531	-633820.4688
229866.9063	489298.1563	1058893.375	4325071	-1023570.063	-227643	-961018.6875	136663.125
1205245.625	2386297.75	4134517.75	4209640	-8267742.375	-9116313.133	-9054034.375	-8748847.125
4838220.5	5989610	7196361	7528256.5	-170502362	-156098350	-144139452	-155251562
680086.625	7122028.5	3757842	4863109.5	-43254391.25	-44739429.88	-42880126	-40628191.5
6614905.5	8474995	4974632.5	12607139	-21855384	-24289794	-20991455	-21518870
1809379	3166296.5	4690755	2627413.25	-55810420	-60561396	-52134966	-56077972
2263918	830624.4375	4329830.5	10571793	-18316043.25	-16365657	-14891428	-15227425.5
-422677.3125	232356.2344	1123875	2746210.5	1260457.25	-1450561.25	-828111.5	101014
393759.875	1359005.75	1329041.125	2745960.25	2505653.137	2954838.031	2476717.504	3066468.438
1965582.375	4272043.5	3442731	4518502	-10066473	-13644846	-16204924.75	-13169884.5
567223.3125	9596649	3692350	7254443.5	-12858188	-19651287	-17660868.5	-15906087
-179255.6563	-3388808	7529950.5	3677222.75	-21820827.5	-21620684	-21853104	-21528893.75
1040165.625	6428225.5	6093796.5	1911645.75	-34843994.5	-35182859.5	-34070036	-34218104

Table 15: Full results of Influence on both tasks for the Made-Up Entity dataset