The signal is coming from inside the noun phrase! Tracking semantic proto-role inferences during sentence processing

Lucas Y. Li
Cornell University
lyl27@cornell.edu

Zander Lynch
Cornell University
zcl7@cornell.edu

Marten van Schijndel
Cornell University
mv443@cornell.edu

Abstract

Semantic roles between a predicate and argument can be decomposed into proto-role properties (e.g., INSTIGATION). We introduce a novel LLM feature attribution method, Generalized Contextual Decomposition for Transformers (GCD-T), which we use to probe which parts of a sentence enable models to infer proto-role properties. We compare our findings with human inferences.

1 Introduction

Semantic roles characterize the relationship between entities and their predicates. For example, agents are initiators of events while patients receive event outcomes. These broad roles can be decomposed into combinations of specific semantic features called *proto-role properties* (Dowty, 1991), which are categorized by whether they reflect more agent-like or patient-like features. When processing a sentence, one must infer these proto-role properties to form an understanding of the events being conveyed (who caused the event and how were others impacted?). However, it remains an open question which parts of a sentence convey these properties to readers.

Various linguistic theories have posited different lexical categories to project proto-role properties but these theories have not been tested against what processors actually use to infer these properties in a sentence. Following Jackendoff (1976) and Davidson (1967), Dowty (1989) saw semantic roles and thus proto-role properties (PRPs) as "determined completely and solely by verb meanings." For example under this view, x *murders* y entails that x is VOLITIONAL no matter the identity of x, while x *kills* y does not, and both entail y is SENTIENT. Other scholars argue that semantic roles are also determined by the semantic properties of the arguments or the syntax (Parsons, 1990; Kratzer, 1996; Williams, 2015; Husband, 2023).

Even if semantic roles *are* fully specified by the verb, arguments may still be required to carry the relevant PRPs in order to be validated against the argument slots entailed by the verb. This explains why (1a) is acceptable while (1c-d) are not and (1b) is only allowed metaphorically, as *the disease* cannot be SENTIENT nor VOLITIONAL.

(1) a. The disease killed Mary.
b. ? The disease murdered Mary.
c. # Mary killed the disease.
d. # Mary murdered the disease.

Intuitively, we know that *Mary* is SENTIENT and *the disease* is not, irrespective of whether this is also lexically entailed by the verb. Therefore, our research question is *how much* of the signal used to predict PRPs comes from the argument compared to the verb, or the other words in the sentence? In this work, we introduce a novel analytical method to probe the contribution of nouns, modifiers, and verbs to PRP prediction in LLMs and validate our findings with an offline judgment task by humans. While we do not assert that LLMs model human cognition, they learn *statistical regularities* over their training corpora and thus can reveal what information is useful for accurate semantic inference. We find that both LLMs and human annotators utilize nouns and adjectives in addition to verbs to infer proto-role properties.[1] We list our major contributions below:

- We implement Generalized Contextual Decomposition for Transformers (GCD-T) to quantify the contributions of arguments, modifiers, and predicates in the Semantic Proto-Role Labelling (SPRL) task.
- We find that NPs are at least as informative for SPRL as verbs, especially for grammatical

[1]The code for our experiments can be found at https://github.com/lucas-y-li/decomposing_transformers

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 427–441
July 3-4, 2026 ©2026 Association for Computational Linguistics

agents (active subjects and passive objects) and agentive properties.

- We confirm experimentally that these computational results generalize to humans. We find that both nouns and adjectives significantly affect human PRP judgments of agentive properties, indicating that humans use non-verb elements when inferring semantic role properties.

2 Background

2.1 Proto-Role Property Labelling

PRP judgments have been experimentally studied using likelihood ratings of questions that accessibly paraphrase each property (Kako, 2006). For example, to elicit a rating for VOLITION, participants were asked, "How likely is it that the [Arg] chose to be involved in the [Pred]?" Reisinger et al. (2015) and White et al. (2016) later scaled these elicitations to produce a large-scale database of crowdsourced annotations, which they released as the Universal Decompositional Semantics dataset (Decomp).[2]

Following the introduction of Decomp, several works have presented computational models for the Semantic Proto-Role Labelling (SPRL) task, which aims to predict the likelihood of each PRP for an argument-predicate pair in a sentence. Teichert et al. (2017) first approached the task as a binary multi-label classification problem using a Conditional Random Field (CRF) model. Similarly, Rudinger et al. (2018) proposed a bidirectional long-short term memory (BiLSTM) model trained with multi-task learning and found that predicting all PRPs jointly resulted in effective parameter sharing. Our work is the first to probe SPRL classifiers to investigate which tokens in the input the models utilize when making decisions.

2.2 Pre-Verbal Semantic Role Prediction

Existing psycholinguistic research has found cross-linguistic evidence that humans use semantic information from pre-verbal NPs to infer semantic roles while reading sentences, prior to seeing the verb (Bornkessel et al., 2003; Chow et al., 2018; Liao et al., 2022). These studies argue that, based on an incremental processing model, humans must make preliminary predictions about the semantic roles of pre-verbal NPs using incomplete information. However, the influence of post-verbal NPs on semantic role inference, when the complete sentence is available to the readers, is not yet known (Husband, 2023). These studies were also limited in that they only tested small sets of controlled stimuli. Our computational approach allows the comparison of event elements in a large corpus of naturalistic sentences.

Property Type	Proto-Role Properties
Proto-Agent	AWARENESS, SENTIENT, VOLITION, INSTIGATION, EXISTEDBEFORE, EXISTEDDURING, EXISTEDAFTER
Proto-Patient	CHANGEOFSTATE, CHANGEOFLOCATION, CHANGEOFPOSSESSION

Table 1: Proto-role properties used in our analysis.[3]

3 Interpretability Analysis Method

3.1 Generalized Contextual Decomposition

To quantify the degree to which the models utilize NPs and verbs during SPRL, we implement Generalized Contextual Decomposition (GCD), an interpretability method originally proposed for LSTM gates (Murdoch et al., 2018; Jumelet et al., 2019), which we further generalize to self-attention to apply it to Transformer-based models.

Given an input sequence $\mathbf{x}$, GCD partitions the tokens into *in-focus* (β) and *out-of-focus* (γ) components, $\mathbf{x}_\beta$ and $\mathbf{x}_\gamma$. This is done through masking, such that $\mathbf{x} = \mathbf{x}_\beta + \mathbf{x}_\gamma$, and each component has zero embeddings for the tokens not assigned to it. The bias component $\mathbf{x}_\delta$ is initialized as zeros. The goal of GCD is to decompose the output of encoder $\mathcal{E}$ into its constituent contributions $L_\mathcal{E}$,

$$\mathcal{E}(\mathbf{x}) = L_\mathcal{E}(\mathbf{x}_\beta) + L_\mathcal{E}(\mathbf{x}_\gamma) + L_\mathcal{E}(\mathbf{x}_\delta) \quad (1)$$

This is accomplished by propagating the in-focus and out-of-focus partitions through all layers of $\mathcal{E}$. For any layer ℓ, its hidden state representation $\mathbf{h}^\ell$ is linearly decomposed into

$$\begin{aligned}\mathbf{h}^\ell &= L_\ell(\mathbf{h}_\beta^{\ell-1}) + L_\ell(\mathbf{h}_\gamma^{\ell-1}) + L_\ell(\mathbf{h}_\delta^{\ell-1}) \\ &= \mathbf{h}_\beta^\ell + \mathbf{h}_\gamma^\ell + \mathbf{h}_\delta^\ell\end{aligned} \quad (2)$$

where $\mathbf{h}_\beta^\ell$ contains the information the layer received from the in-focus input, $\mathbf{h}_\gamma^\ell$ contains information from the out-of-focus input, and $\mathbf{h}_\delta^\ell$ contains information from the bias parameters of layers $\leq \ell$. Thus, GCD reveals how much, and whether positively or negatively, each component *contributes to* each classifier logit. This is distinct from model interpretability methods that analyze

[2] https://decomp.io/projects/semantic-proto-roles

[3] CHANGEOFLOCATION is sometimes considered a proto-agent property. See §4.4 for more discussion of why we consider it a proto-patient property in the Decomp dataset.

attention weights, which instead reveal the information a model has *access to* for classification.

The output of a fully-connected layer can be easily linearly decomposed,

$$L_{FC}(\mathbf{x}_i) = \mathbf{W}^T\mathbf{x}_i, \;\; i \in \{\beta, \gamma\}$$
$$L_{FC}(\mathbf{x}_\delta) = \mathbf{W}^T\mathbf{x}_\delta + \mathbf{b} \quad (3)$$

Non-linear activation functions are decomposed using Shapley values (Shapley, 1953).

3.2 GCD for Transformers

In this work, we introduce Generalized Contextual Decomposition for Transformers (GCD-T). We follow Jumelet et al. (2019) in allowing out-of-focus material to inform decisions about *when* to use in-focus information (and vice-versa), but the representations themselves are kept in isolated in-focus and out-of-focus partitions.[4]

Vaswani et al. (2017) defines self-attention as

$$Attn(\mathbf{x}) = \mathbb{A}(\mathbf{x})\mathbf{V}, \;\; \mathbb{A}(\mathbf{x}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}),$$
$$\mathbf{Q} = \mathbf{W}_Q^T\mathbf{x}, \; \mathbf{K} = \mathbf{W}_K^T\mathbf{x}, \; \mathbf{V} = \mathbf{W}_V^T\mathbf{x} \quad (4)$$

Where $\mathbb{A}(\mathbf{x})$ is the attention weights, and $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the queries, keys, and values, respectively. We do not decompose $\mathbb{A}(\mathbf{x})$ so the layer receives the full context to determine which $\mathbf{x}_\beta$ values to propagate forward.$\mathbf{V}$ is decomposed linearly as normal. Thus, the decomposition is

$$L_{Attn}(\mathbf{x}_i) = \mathbb{A}(\textstyle\sum_i \mathbf{x}_i)\mathbf{W}_V^T\mathbf{x}_i, \; i \in \{\beta, \gamma, \delta\} \quad (5)$$

Ultimately, we want to determine the contributions of the in-focus partition to the classification logits ($\mathbf{z}_\beta$). While the original Murdoch et al. (2018) CD method used raw output logits ($\hat{\mathbf{z}}_\beta$), we normalize the raw outputs by the token length of the in-focus element ($|\beta|$): $\mathbf{z}_\beta = \hat{\mathbf{z}}_\beta / |\beta|$. These per-token contributions prevent long β spans (such as complex NPs) from having disproportionately large contributions.[5]

4 Semantic Proto-Role Inference

We now turn to our main research question: investigating which tokens give rise to PRP inferences during processing.

[4]This differs from Murdoch et al., who fully partition representations across all layers. Further discussion about the differences between GCD and CD and an empirical analysis with respect to Transformers is given in §B.

[5]An empirical validation of GCD-T by comparing behaviour to GCD on the tasks of subject-verb number agreement and sentiment analysis is shown in §C.

4.1 Dataset

Following prior work on SPRL, we use the Decomp dataset (Reisinger et al., 2015; White et al., 2016). The dataset consists of 15,000 PRP judgments across 10 properties, shown in Table 1. Annotators used a 5-point likert scale to judge the appropriateness of a PRP applying to a sentence (see §2.1). As is standard with this data, we binarized the judgments into ≤ 3 and > 3. Full preprocessing details and examples are given in §A.

4.2 Models

We evaluate the Transformer models RoBERTa-large (355M parameters; Liu et al., 2019) and GPT-2-medium (345M parameters; Radford et al., 2019). The models are the same size, however RoBERTa is encoder-only and has bidirectional context while GPT-2 is autoregressive and thus must base its decisions only on left context, analogously to humans during incremental processing.[6] We fine-tune the LLMs for the binary semantic proto-role labelling (SPRL) task on Decomp.[7] We insert custom special tokens <|*pred*|> and <|*arg*|> around the relevant predicate and argument of each sentence to indicate which argument-predicate pair is being queried.

Evaluation shows that both models perform comparably to average human inter-annotator agreement on all properties (Table 2).[8] There is no notable difference in F1 scores between active and passive sentences nor between models. Although GCD-T can be applied to any large Transformer LLM with trivial modifications, we choose to probe RoBERTa-large and GPT-2-medium as they already perform comparably to humans on the SPRL task despite their small size.

4.3 Experimental Setup

We partition each sentence-argument-predicate triplet into the following in-focus token spans: ARG, which contains the tokens of the target argument excluding any modifiers; MOD, the modifier tokens of the target argument; PRED, the tokens of the predicate verb; and OTHER, which consists of all non-target arguments of the predicate, if any.

[6]We also experimented with the smaller models RoBERTa-base, GPT-2-small, DistilGPT, and DistilRoBERTa (Sanh et al., 2019) and found similar results.

[7]Both models were fine-tuned with hyperparameters {*lr=5e-5, λ=0.1, epochs=4, batch_size=16, dropout=0.1*}.

[8]Each Decomp V2 item has two annotator ratings, so human F1 is calculated as pairwise micro-F1. Model F1 is calculated as average micro-F1 with respect to both annotations.

Property	Humans		RoBERTa		GPT-2	
	ACT	PASS	ACT	PASS	ACT	PASS
awareness	89.7	88.9	**91.7**	89.7	91.2	**90.5**
Δ_location	77.9	**77.8**	**80.0**	77.4	81.8	79.0
Δ_state	65.0	**68.3**	**67.2**	65.5	**69.3**	68.3
Δ_possession	91.3	90.1	**92.8**	**91.1**	92.5	90.7
existed_after	85.5	86.9	**87.5**	**88.3**	87.2	87.5
existed_before	85.0	85.7	**86.7**	**86.9**	86.4	84.1
existed_during	96.1	96.4	**97.2**	**97.4**	96.7	97.8
instigation	72.4	69.0	**75.7**	**71.4**	73.0	68.3
sentient	89.5	88.5	**92.1**	**90.3**	92.6	91.5
volition	86.8	88.1	**88.2**	**88.5**	86.3	87.7
average	83.9	84.0	**85.9**	**84.6**	85.7	84.5

Table 2: Decomp inter-annotator and model F1.

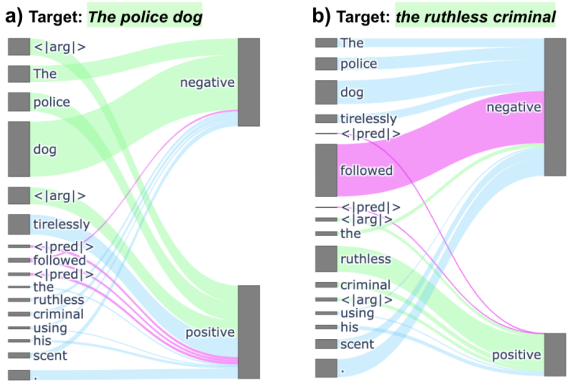


Figure 1: Example SPRL GCD-T GPT-2 token contributions for INSTIGATION. Target NP (MOD, ARG) contributions are shown in green and PRED contributions are in pink.

Argument modifiers are obtained using the gold-label dependency relations and syntactic parses from Universal Dependencies (Silveira et al., 2014). We recursively consider any optional adjunct and the nodes it dominates to be modifiers, unless the argument itself is an adjunct (e.g., PP), in which case we start with its children.

GCD-T is applied to obtain each component's contribution to each PRP. We first show an example decomposition before presenting a quantitative analysis of our Decomp V2 test set.

4.4 GCD-T Results

Qualitative Example. Figure 1 shows an example GPT-2 decomposition for the proto-agent property INSTIGATION of the sentence "The police dog tirelessly followed the ruthless criminal using his scent." In Figure 1a, the target argument is *The police dog* and in Figure 1b, the target argument is *the ruthless criminal*. Both arguments share the *followed* predicate. Based on our categorization, *ruthless* is considered a modifier, but *police* is not because *police dog* is a compound noun.

In Figure 1a, GPT-2 identifies *dogs* as typically non-instigative, which is correct given the training data as non-human living entities are less likely to initiate actions than humans. However, the attribute *police* is identified as contributing positively to instigation. In Figure 1b, the model identifies objects of *followed* as typically non-instigative and assigns it a negative contribution, while the modifier *ruthless* contributes positively.

Quantitative Results. First, we observe that although Reisinger et al. (2015) described CHANGEOFLOCATION as a proto-agent property, GCD-T reveals that it behaves more similarly to proto-patient properties[9] and we thus consider it as such. This categorization is supported by Dowty (1991), who states that CHANGEOFLOCATION is agentive only if the other argument is non-instigative, otherwise it is patientive. As annotators labeled each PRP in Decomp separately, they would not have been aware of this detail, so in this way we can see that GCD-T actually provides empirical evidence for a theoretical PRP distinction.

For proto-agent properties (Figure 2a, 2c), the ARG rarely contributes negative evidence, and only contributes positive evidence in active subjects or passive objects (grammatical Agents). In RoBERTa, ARG and PRED contribute equally (Figure 2a), however GPT-2 disproportionately identifies agentiveness using the ARG (Figure 2c). Surprisingly, passivization has only a small effect on ARG contributions, as contribution trends are largely the same for active subjects and passive objects (grammatical Agents), and active objects and passive subjects (grammatical Patients). Therefore, it appears there is little difference in how models use the PRP cues of post-verbal NPs compared with pre-verbal NPs, even for the incremental model GPT-2.

For proto-patient properties, ARG and PRED contribute largely negatively in RoBERTa (Figure 2b) but positively in GPT-2 (Figure 2d). Relative to ARG, PRED contributes more to GPT-2's proto-patient interpretation than its proto-agent interpretation, particularly in active sentences, indicating that agentive inferences primarily originate from arguments, while patienthood primarily originates from verbs with some input from entities. This is consistent with theoretical proposals that agents,

[9] CHANGEOFLOCATION contributions have mean KL divergence of 6.9 from proto-agent properties and 2.5 from proto-patient properties in RoBERTa and 5.7 from proto-agent properties and 4.5 from proto-patient properties in GPT-2.

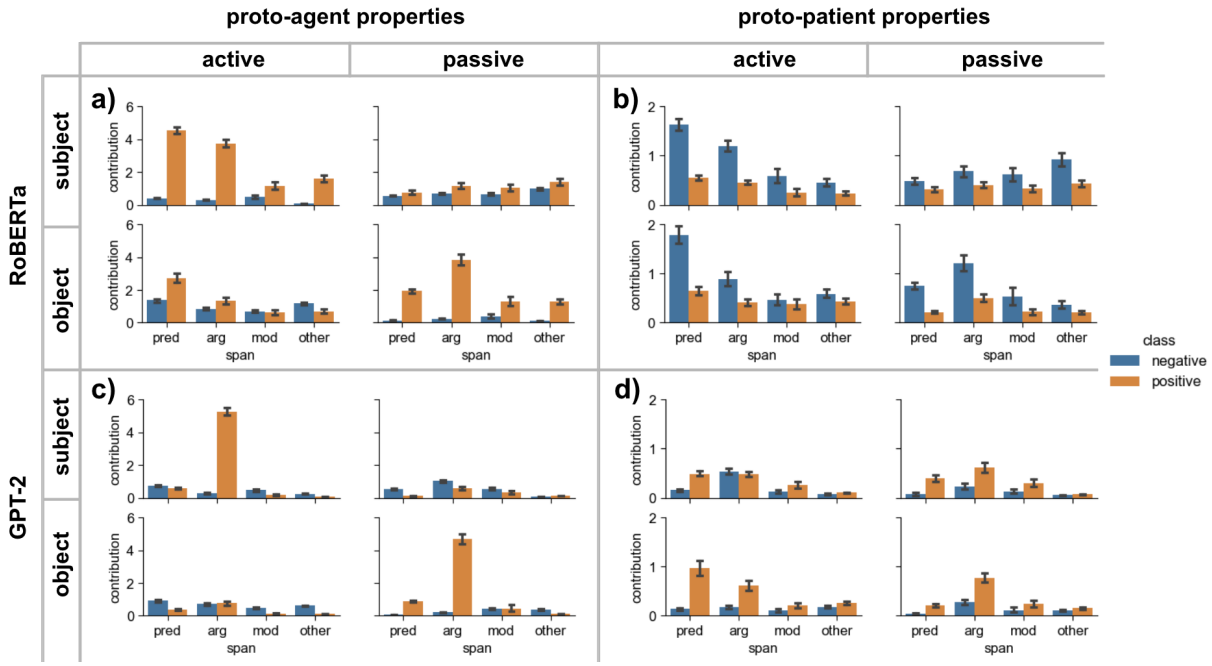


Figure 2: Average GCD-T SPRL token contributions of event constituents. Scales differ between proto-agent and proto-patient contributions likely because of the greater frequency of proto-agent properties in the training data (e.g., in intransitive sentences). Error bars represent 95% CI.

but not patients, are severed from the verb as the verb combines directly with the latter and not the former (e.g., Kratzer, 1996; Levin, 2012).

While GPT-2 identifies positive evidence for both proto-agent and proto-patient properties (Figure 2c, 2d), RoBERTa focuses mostly on positive evidence of *agentiveness* even when the goal of the task is to identify a proto-patient property (Figure 2a, 2b). This perhaps indicates the strength of proto-agentive features learned by RoBERTa during pretraining. Furthermore, although MOD and OTHER contribute less to SPRL than ARG and MOD, their contributions are still non-zero. This is particularly interesting for OTHER, as it means semantic role properties can depend on other sentence elements *outside* of the target predicate-argument pair. Overall, it is clear that both models strongly utilize non-predicate elements for SPRL of all PRPs and syntactic positions.

4.5 Contextualization Analysis

A possible algorithmic hypothesis about what causes the above results is that, rather than relying on ARG cues to predict PRPs (as the above results suggest), the models could rely on ARG to identify the word sense of PRED, and once the correct sense is selected, the models use that verb sense to infer the PRPs. Because GCD-T isolates ARG from PRED, the results look like the models use ARG because that is the first step in their processing pipeline, when in fact they may rely on the sense-disambiguated PRED in practice. To test this hypothesis, we move the GCD-T partition layer to later model layers. If the models use ARG to contextualize PRED, then we should see PRED contributions increase in later layers.

Method. We compute the SPRL contributions but vary the partitioning layer (the point at which in-focus and out-of-focus elements are partitioned from one another)[10] over the 24 Transformer blocks of each model. The average percent-contribution is measured as a proportion of all event elements; that is, the percent-contributions of ARG, MOD, PRED, and OTHER sum to 100% in each layer.

Results. In the percent-contributions of RoBERTa (Figure 3a), the ARG and PRED embeddings contain approximately equal amounts of thematic information for the first 20 Transformer blocks. In the last few layers, however, the importance shifts towards ARG and MOD while the relative contributions of PRED and OTHER decrease. In GPT-2 (Figure 3b), the relative contribution of ARG gradually decreases as those of PRED and OTHER increase. As with RoBERTa, the direction of the trends switch after approximately layer 20.

These results suggest that GPT-2 may adhere to something like the hypothesized processing

[10] See §C.2 for more information about partitioning layers.

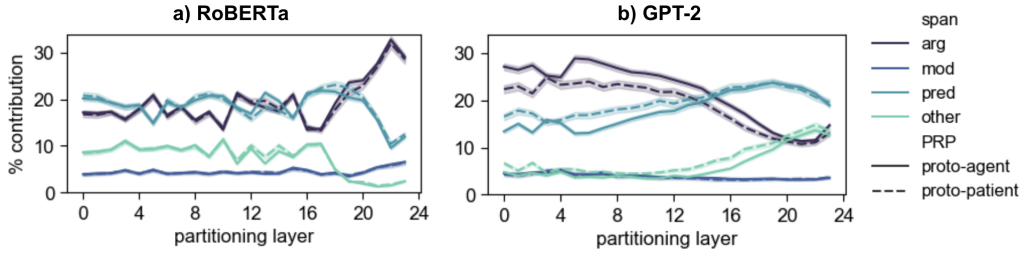


Figure 3: Average proportional GCD-T token contribution by partitioning layer. Ribbons represent 95% CI.

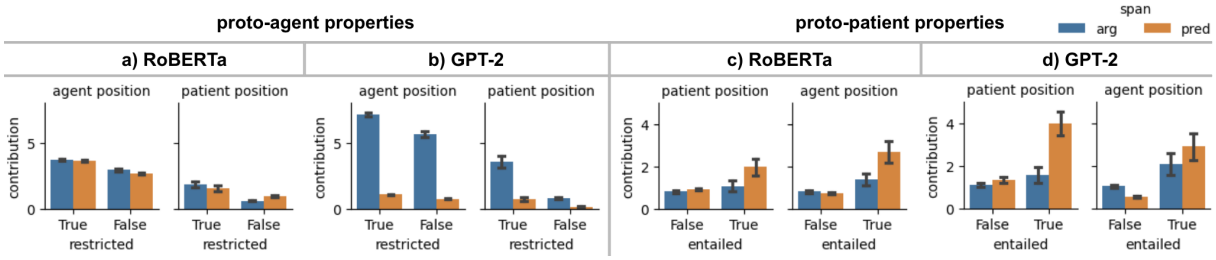


Figure 4: GCD-T SPRL positive token contributions by verbal lexical specification. Restricted denotes that the PRP is a selectional restriction and entailed denotes it is a lexical entailment. Agent-position denotes active subjects/passive objects; patient-position denotes active objects/passive subjects. Error bars represent 95% CI; note that errors are higher for proto-patient properties due to smaller sample sizes.

pipeline: initially focusing primarily on ARG, which is then used to reinterpret the semantic information from PRED. RoBERTa seems to do the opposite, contextualizing ARG using PRED. Future studies could explore which conditions give rise to different contextualization strategies, but this is beyond the scope of this paper.

5 Human Experiments

5.1 Research Questions and Design

Our GCD-T analysis revealed that LLMs rely on non-verb sentential elements to make PRP judgments, even for post-verbal arguments, and that this was especially true for proto-agentive properties. In this section we test whether this finding holds for humans as well. To that end, we elicited PRP judgments from humans using stimuli with various elements masked to see how masking of sentence components influenced their proto-agentive PRP inferences in pre-verbal and post-verbal contexts.

To elicit PRP judgments, we adopt the annotation paradigm of Reisinger et al. (2015) and Kako (2006). However, to isolate the contribution of each component (PRED, ARG, and MOD), we replace the other two regions with phonologically-plausible pseudo-words in the style of Berko (1958). In doing so, participants must use the semantic content of the remaining real word to make their PRP judgments. This allows us to assess the extent to which inferences are made on the basis of cues present in each element. Since our GCD-T analysis suggested that proto-agentive PRPs should be most influenced by non-verbal elements, we focused on the following proto-agentive PRPs highly correlated with animacy: `A`, `I`, `S`, and `V`.

Participants. We recruited 62 participants from the Prolific crowdsourcing platform,[11] all located in the United States and self-reported monolingual English speakers. Participants were compensated at a rate equivalent to $15/hour.

5.2 Experimental Setup

Materials. There were four levels of the real word factor (PRED, MOD, INANIMATE-ARG, and ANIMATE-ARG), which varied which of the potential PRP triggers was left as a real word while others were masked by pseudo-words. There was also a position factor with two levels, Subject and Object, which varied whether the argument of interest was the grammatical subject or object of an active transitive sentence allowing us to vary the target argument's syntactic position. These factors were fully crossed resulting in 8 total conditions. Stimulus sentences were all of the form "[Proper Name] [PRED]-ed the [MOD] [ARG]" or "The [MOD] [ARG] [PRED]-ed [Proper Name]".

We use state-of-mind adjectives for MOD as we hypothesize they are the most likely to increase animacy. However, all sentences include an adjective to control for any focus or salience effect result-

[11] https://www.prolific.com

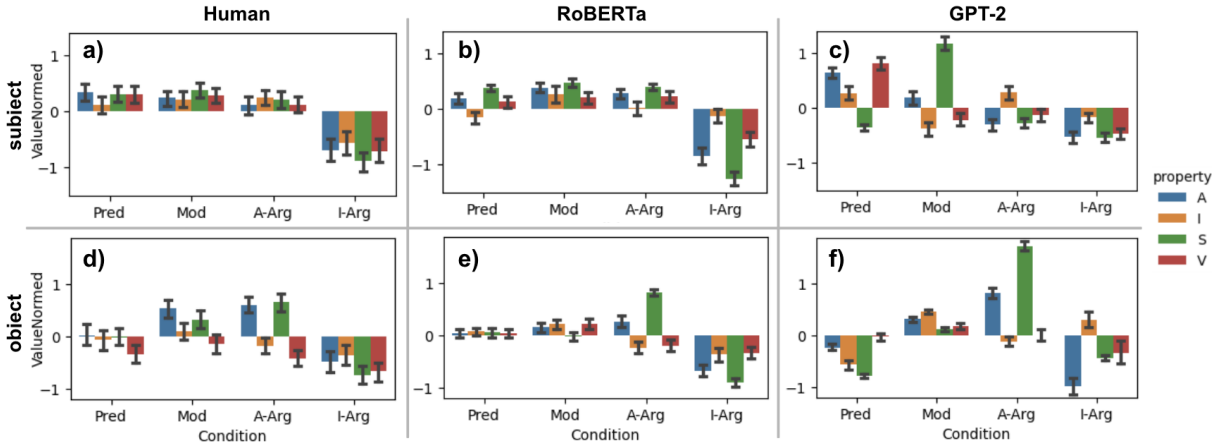


Figure 5: Average SPRL human ratings and model GCD-T contributions, z-scored by PRP and syntactic position. A-Arg denotes ANIMATE-ARG and I-Arg denotes INANIMATE-ARG. Error bars represent 95% CI.

ing from modifying an argument at all. Similarly, when the ARG is unmasked, we use two classes of nouns, Animate (animals) and Inanimate. The real PREDs were selected from a list of the most frequent active transitive verbs in Decomp V1. To study the influence of selectional restriction, we chose verbs which, according to VerbNet, require animate subjects, but may use both animate and inanimate objects.

We create 16 sets of stimuli, each with a unique real proper name, verb, adjective, inanimate noun, and animate noun. Example stimuli are shown in §E. Pseudo-words were generated using the Python package `gibberish`[12] and compared to transition probabilities for English orthography to ensure phonological plausibility. We also manually checked the generated pseudo-words to verify they do not look like misspellings of real English words nor can be pronounced as real English words. Proper names were gender-balanced.

Procedure. Prior to the experiment, participants underwent a training phase[13] and were told that they would be shown unfamiliar words and would need to use the context to figure out the meaning of each sentence. Participants saw only one of the eight sentence arrangements of each item, in addition to seeing only Subjects or only Objects, to reduce confusion and minimize potential confounding factors. They were asked to rate the likelihood of all four PRPs on each sentence however each sentence-PRP question pair was displayed on a different screen and randomized throughout the trials to reduce cross-contamination between PRP questions. As such, each participant saw 64 total experimental trials.

Models. In addition to collecting human judgments, we also obtained GCD-T contributions from RoBERTa and GPT-2 following the methods described in §4. To align with humans who give a single PRP rating, we sum all GCD-T logits, positive and negative, to get a single PRP rating.

To decrease the variance of the model results, we augmented the stimuli by generating all combinations of argument-predicate pairs, resulting in 256 sets. Unlike with humans, it is not problematic for the LLMs to see all arrangements of each set, as their parameters are frozen after fine-tuning on Decomp. As GCD-T allows for the contribution of each component in a sentence to be computed independently in a single forward pass, the use of pseudo-words is not necessary to probe the models. Thus, the models receive two sentences from each set, one with the ANIMATE-ARG and one with the INANIMATE-ARG, and both containing the PRED and MOD. PRED and MOD contributions are averaged across both sentences when the target argument is the object, but computed using only the ANIMATE-ARG with subject targets.[14]

5.3 Results and Discussion

Figure 5 shows the average contributions for the human annotators and LLMs, z-scored by syntactic position and PRP. RoBERTa accurately captures word-level human ratings in all conditions (Figure 5b, 5e), despite only being trained on

[12] https://github.com/greghaskins/gibberish

[13] The practice questions during the training phase were completely separate from the experiment and given only to familiarize participants with the 5-point Likert scale.

[14] This is because sentences with the INANIMATE-ARG subjects were ungrammatical due to violating the semantic restrictions of the PRED. Indeed, humans participants never see the true PRED and INANIMATE-ARG together, and likely assume the pseudo-word subject represents an animate noun.

annotations using full-sentence contexts. Of the tested properties, RoBERTa struggles most with I, however this property also has the lowest inter-annotator agreement in humans. Conversely, GPT-2 fails to predict human ratings, particularly in subjects (Figure 5c). This result makes sense since both RoBERTa and the humans have access to the full sentence context when making their judgments, while GPT-2 does not.

All analyses of the human data are done using the ANIMATE-ARG condition as the baseline for the comparison. We see that in the object syntactic position (Figure 5d), A and S judgments are significantly lower in PRED and INANIMATE-ARG compared to ANIMATE-ARG. The difference in S between MOD and ANIMATE-ARG is smaller but still significant. Additionally, the MOD condition shows moderately significant increases in V and I while INANIMATE-ARG showed a moderately significant decrease in V. These results show that MOD, ANIMATE-ARG, and INANIMATE-ARG have larger impacts on PRP inferences than PRED, which is notable given the centrality of predicates in conveying events. We will return to this in the General Discussion.

In the subject position (Figure 5a), S and V for the MOD condition and V and A for the PRED condition are significantly higher than the corresponding values for ANIMATE-ARG. These results suggest that, aside from S, active-voice subjects are agentive by default (as most lexical items in subject position don't influence PRPs) but can be manipulated by verbs and adjectives.

These human findings concur with our GCD-T results that non-verbal elements contribute to PRP inferences in readers. We even see, through the different conditions, that these various grammatical elements contribute to PRPs in distinct ways. Most notably, we see that MOD and ANIMATE-ARG increase A and S, especially in the object condition, while the INANIMATE-ARG condition decreases the inferences of all four properties.

6 General Discussion

We introduced a new method, Generalized Contextual Decomposition for Transformers (GCD-T), for feature attribution in LLMs, based on previous (Generalized) Contextual Decomposition methods for earlier neural architectures (Murdoch et al., 2018; Jumelet et al., 2019). GCD-T requires no additional training or gradient computation and is therefore less expensive than other probing methods. One common concern with LLM probing methods is that the probes used to interpret an LLM can produce false positives because the probes can learn to solve the probe tasks even when the LLM does not (Hewitt and Liang, 2019). GCD-T avoids this issue because it simply decomposes the representations within a model to disentangle the input-output mappings learned by the model during training.

We used GCD-T to investigate how different event elements contribute to the inference of semantic proto-role properties (PRPs) in Transformer language models. *A priori*, one could imagine several different ways that processors could infer PRPs. First, because events are communicated using predicates, one could imagine that processors would rely as much as possible on predicates, minimizing the ability of arguments to modulate PRP inferences. We find that is not the case, and that LLMs rely heavily on arguments and modifiers to infer PRPs. Second, readers could adopt a maximally incremental approach where they use pre-verbal arguments to infer PRPs but they neglect to use post-verbal arguments, having already seen the verb and much of the sentence. We show that they do not. Post-verbal material is used similarly to pre-verbal material. Finally, all PRPs could be inferred using the same procedure (e.g., all PRP inferences could rely 20% on the argument and 80% on the predicate). We find that this is not the case either: proto-agent properties rely more on argument cues while proto-patient properties rely more on predicate cues, though we find differences within these broad proto-role categories as well.

Finally, we validated the results of our computational experiments using a human experiment. We confirmed that humans also heavily use non-predicate elements to infer PRPs, even in post-verbal sentence contexts. In fact, we found that non-verbal elements had *larger* impacts on PRP inferences than predicates in our task. As with LLMs, each PRP was affected differently by our experimental manipulations, indicating that readers adopt different strategies for inferring each PRP. There remains much to do in this area, including analyzing more proto-role properties, and exploring the specific ways in which verbal selectional restriction and lexical entailments influence each PRP inference in humans, but we hope our work can serve as a first step in this direction.

References

Jean Berko. 1958. The child's learning of english morphology. *WORD*, 14:150–177.

Ina Bornkessel, Matthias Schlesewsky, and Angela D Friederici. 2003. Eliciting thematic reanalysis effects: The role of syntax-independent information during parsing. *Language and Cognitive Processes*, 18(3):269–298.

Wing-Yee Chow, Ellen Lau, Suiping Wang, and Colin Phillips. 2018. Wait a second! delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7):803–828.

Donald Davidson. 1967. Truth and meaning. In *Philosophy, Language, and Artificial Intelligence: Resources for Processing Natural Language*, pages 93–111. Springer.

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.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

David R. Dowty. 1989. On the Semantic Content of the Notion of 'Thematic Role'. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types and Meaning: Volume II: Semantic Issues*, pages 69–129. Springer Netherlands, Dordrecht.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

E. Matthew Husband. 2023. Thematic separation in light of sentence comprehension. *Language and Linguistics Compass*, 17(4):e12496.

Ray Jackendoff. 1976. Toward an Explanatory Semantic Representation. *Linguistic Inquiry*, 7(1):89–150.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment.

Edward Kako. 2006. Thematic role properties of subjects and objects. *Cognition*, 101(1):1–42.

Angelika Kratzer. 1996. Severing the external argument from its verb. *Phrase structure and the lexicon/Kluwer*.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Beth Levin. 2012. On Dowty's "Thematic proto-roles and argument selection". In *A Reader's Guide to Classic Papers in Formal Semantics: Volume 100 of Studies in Linguistics and Philosophy*, pages 103–119. Springer.

Chia-Hsuan Liao, Ellen Lau, and Wing-Yee Chow. 2022. Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language*, 126:104350.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *International Conference on Learning Representations*.

Terence Parsons. 1990. *Events in the Semantics of English: A study in subatomic semantics*. MIT Press.

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.

Dee Ann Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-Davidsonian Semantic Proto-role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew Gormley. 2017. Semantic Proto-Role Labeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Aaron Steven White, Dee Ann Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Alexander Williams. 2015. *Arguments in Syntax and Semantics*. Cambridge University Press.

A Dataset Preprocessing And Examples

In this section, we describe in greater detail the datasets used for the SPRL task, Decomp V1 (Reisinger et al., 2015) and Decomp V2 (White et al., 2016), in addition to our preprocessing methods. Examples from the Decomp dataset for SPRL are shown in Table 3.

Decomp V1 contains PRP judgments of 9,738 predicate-argument pairs in the Penn Treebank WSJ corpus (Marcus et al., 1993). Annotators rated the likelihood of each pair and PRP on a 5-point Likert scale (1=*very unlikely*, 2=*somewhat unlikely*, 3=*not enough information*, 4=*somewhat likely*, and 5=*very likely*). Negated and modal verbs were removed.

Decomp V2 contains PRP judgments of 6,091 predicate-argument pairs from the Universal Dependencies English Web Treebank (Silveira et al., 2014). The same annotation protocol was used as V1, with some modifications, including switching from using one annotator to two for each pair.

In our combined dataset, we use the 10 PRPs annotated in both V1 and V2 (see Table 1). We use binned scores of > 3 and ≤ 3 following Teichert et al. (2017) and Rudinger et al. (2018). This corrects for class imbalances, as only 8.8% of pre-binarized scores are 2 or 4. As more than 96% of predicates are in active voice, we deterministically passivized active transitive verbs from the corpus to preserve semantic roles but balance syntactic positions. We manually checked augmented sentences to ensure grammaticality. We assume that passivization generally does not affect PRP ratings and so use the active PRP ratings for our generated passives; while passivization can result in decreased agentiveness (e.g., $5 \rightarrow 4$), we believe it is unlikely to cause a PRP rating to flip from ≤ 3 to > 3, or vice-versa. To offset the large number of active, non-transitive sentences, we upsampled the resulting passive sentences during training to ensure that the models saw a balanced number of passive and active structures. Dataset statistics are shown in Table 4.

We train on Decomp V1 and V2, but the results in Table 2 include only evaluation on Decomp V2 so that inter-annotator agreement can be used as a benchmark against which to compare LLM performance. Evaluation results on the Decomp V1 dataset in comparison to prior models are presented in §D, Table 6.

B Comparison of GCD and CD in Transformers

In this section, we compare the differences between the formulations of GCD (Jumelet et al., 2019) and CD (Murdoch et al., 2018) and explain why we chose to follow GCD for the decomposition of Transformers. Additionally, we provide empirical evidence using the subject-verb number agreement task that GCD provides more intuitive results when used to interpret the behaviour of Transformers than CD.

B.1 Theoretical Motivation for GCD

Any contextual decomposition method needs to decide how to allow the in-focus partition to interact with the out-of-focus partition. On the one hand, keeping the two partitions totally separate simplifies the interpretability objective. On the other hand, the interaction between the in-focus and out-of-focus elements could be critical to the model behavior at hand. The original Contextual Decomposition (CD) method (Murdoch et al., 2018) maximized the separation between the partitions by only considering representations in-focus when they solely consisted of in-focus interactions (rendering out-of-focus any representations that resulted from interactions with out-of-focus material). Formally, the hidden state at time t, $\mathbf{h}_\beta^{\ell,t}$, only consisted of information resulting from the interactions between $\mathbf{h}_\beta^{\ell,t-1}$ embeddings in the previous recurrent cell. In contrast, (Jumelet et al., 2019) argued that this tight in-focus constraint led to a loss of information, such as syntactic knowledge about *when* information should be expressed. Therefore, they allowed, for example, interactions between previous cell state values $\mathbf{c}_\beta^{\ell,t-1}$ and forget gate values $\mathbf{f}_\gamma^{\ell,t}$ to be included in current cell state $\mathbf{c}_\beta^{\ell,t}$, as $\mathbf{f}_\gamma^{\ell,t}$ is only used in the cell state to inform *which* of the $\mathbf{c}_\beta^{\ell,t-1}$ values should be passed on, and is not able to contribute other information.

Thus, following the reasoning of Jumelet et al., we chose not to decompose the attention weights in GCD-T to also allow for interactions between in-focus and out-of-focus queries and keys, so that both partitions may inform which of the in-focus values should be passed onto the next in-focus encodings. Otherwise, we keep the in-focus and out-of-focus partitions separate by decomposing the value vectors. However, we also confirm this decision experimentally.

Sentence	Predicate	Argument	I	V	A	S	EB	ED	EA	CoS	CoP	CoL
I e-mailed your assistant earlier this morning and have had no response .	e-mailed	I	5	5	5	5	5	5	5	1	1	3
		your assistant	1	1	2	5	5	5	5	1	1	1
In the meantime, we extend our best wishes for a safe and happy holiday season .	extend	we	5	5	5	5	5	5	5	1	1	3
		our best wishes for a safe and happy holiday season	1	1	1	1	3	5	5	1	1	1

Table 3: Examples of items in the Decomp dataset for the SPRL task. PRP judgements are shown on the 5-point Likert scale. PRP labels are, from left to right: INSTIGATION, VOLITION, AWARENESS, SENTIENT, EXISTEDBEFORE, EXISTEDDURING, EXISTEDAFTER, CHANGEOFSTATE, CHANGEOFPOSSESSION, and CHANGEOFLOCATION.

Dataset Split	Decomp V1				Decomp V2			
	Active		Passive		Active		Passive	
	Subject	Object	Subject	Object	Subject	Object	Subject	Object
Train	4097	3662	1520	1544	3898	1627	1864	1514
Validation	490	477	183	182	507	231	277	221
Test	477	488	200	177	517	249	263	241

Table 4: Distribution of annotated argument-predicate pairs in Decomp V1 and V2 dataset after passivized sentences were generated. Note that the models were trained and validated on both Decomp V1 and V2, but tested on only Decomp V2 for our analyses. The Object column includes both direct and indirect objects.

B.2 Comparison of GCD-T and CD-T on Number Agreement

We experimented with fully decomposing attention weights based on the method of CD (Murdoch et al., 2018), which we call Contextual Decomposition for Transformers (CD-T). CD-T attention weights are decomposed based on CD's (Murdoch et al., 2018) $f_t \odot c_{t-1}$ and $i_t \odot g_t$ decompositions, such that the β attention weights contain only information from the β and δ $\mathbf{Q}$ and $\mathbf{K}$.

We compare the results of CD-T to GCD-T on the Number Agreement task (see §C.1) in Figure 6. Results show that CD-T failed to accurately capture model behaviour. In BERT, CD-T incorrectly states that singular nouns provide evidence *against* predicting singular verbs, and in GPT-2, it states that plural nouns provide evidence against predicting plural verbs. GCD-T has the expected behaviour and shows that singular nouns provide evidence *for* predicting singular verbs and plural nouns provide evidence for predicting plural verbs in both models.

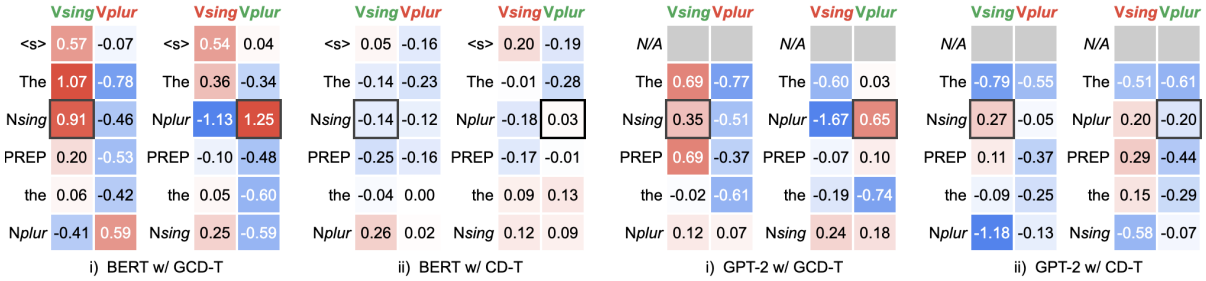
C Validating GCD-T

As GCD-T is a new method, prior to using it in our novel task of studying the locus of signal for proto-role properties, we choose to validate that the method works as intended on two other well-studied interpretability tasks: 1) subject-verb number agreement, which has been widely used in the interpretability literature and was used to motivate the original GCD work, and 2) sentiment analysis, which provides an indication of how the method works with richer semantic signals.

C.1 Number Agreement

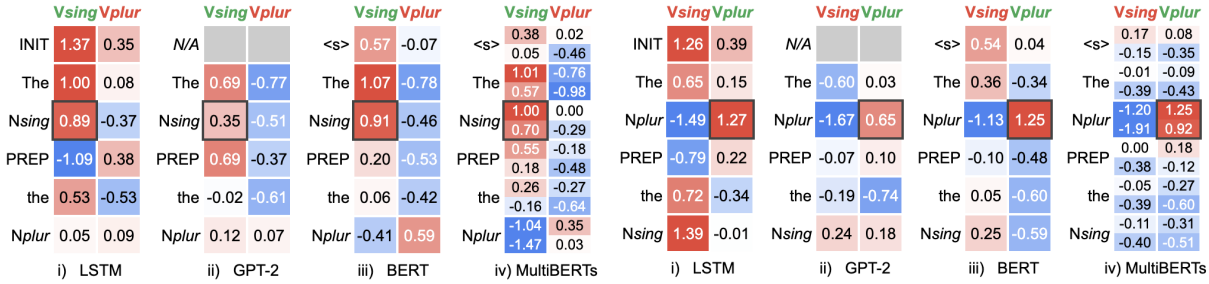
We validate GCD-T against GCD for LSTMs using a Number Agreement dataset (Lakretz et al., 2019) which tasks models with predicting the proper verb form when the sentence contains an intervening distractor of the opposite number. For example, given the sentence prefix "The boy near the cats ..." a model should assign higher probability to the singular continuation 'knows' than the plural 'know' due to the singular subject 'boy'.

We compute average token contributions of pretrained GPT-2-Medium (Radford et al., 2019) and BERT-base-uncased (Devlin et al., 2019) models on this dataset. We discard sentences with multi-token verbs to permit verb masking and we normalize contributions over all single-token verbs in



(a) BERT comparison (b) GPT-2 comparison

Figure 6: Average token contributions using GPT-T and CD-T for the number agreement task. The two columns in each group correspond to GCD-T/CD-T logits over singular and plural verb forms. Red cells indicate positive contributions of that row's token to the verb form tracked by that column (color scaled from [0,1]), while blue cells correspond to negative contributions (color scaled from [-1,0]).



(a) singular subject, plural attractor (b) plural subject, singular attractor

Figure 7: Average token contributions for the number agreement task. The two columns in each group correspond to GCD-T logits over singular and plural verb forms. Black borders indicate the subject's correct number. Red cells indicate positive contributions of that row's token to the verb form tracked by that column (color scaled from [0,1]), while blue cells correspond to negative contributions (color scaled from [-1,0]). INIT denotes initial LSTM state. PREP denotes preposition. MultiBERT values indicate upper and lower bound of the 95% CI across BERT 25 models.

the vocabulary,[15] using a complete verb list obtained from Google Books Ngram. To verify robustness, we also evaluate all MultiBERTs, a set of 25 BERT-base-uncased models pretrained from randomly initialized seeds (Sellam et al., 2022).

Results are presented in Figure 7 with the LSTM results of Gulordava et al. (2018) after reimplementing the GCD method of Jumelet et al. (2019).[16] As expected, GCD-T shows that in both BERT and GPT-2, singular nouns provide positive evidence in predicting singular verbs and plural nouns provide positive evidence for plural verbs. This pattern aligns with the LSTM contributions and demonstrates that the proposed GCD-T is comparable to the original GCD for LSTMs.

C.2 Sentiment Detection

GCD-T also provides insight into how word tokens become contextualized across the layers of the model, by permitting manipulation of the Transformer layer at which tokens are first partitioned into β and γ components. We call this the *partitioning layer*. By default, partitioning occurs prior to the first layer (layer 0), but withholding this until a higher layer allows for inter-component interactions to occur up to the partitioning layer. This application of GCD is an innovation of our work. To illustrate how contextualization analysis can be performed with GCD-T, we apply it to the widely used task of sentiment detection.

We additionally compare GCD-T to traditional Shapley values using the package `shap`.[17] The latter method computes model-level rather than

[15] Normalizing over verbs makes results more interpretable than prior methods since a value of 0 can then be interpreted as the token's logit contribution towards the average verb.

[16] Our LSTM numbers differ slightly from those reported in Jumelet et al. (2019) because we use our per-token contribution measure in our reimplementation of their code. We find the contribution measurements of Jumelet et al., $\mathbf{z}_\beta = \hat{\mathbf{z}}_\beta / \hat{\mathbf{z}} = \hat{\mathbf{z}}_\beta / (\hat{\mathbf{z}}_\beta + \hat{\mathbf{z}}_\gamma + \hat{\mathbf{z}}_\delta)$, are problematic when it is possible for the logits to be negative. For example, if $\hat{\mathbf{z}}$ and $\hat{\mathbf{z}}_\beta$ are both negative, then $\mathbf{z}_\beta > 0$, which is unintuitive given the in-focus partition contributed negatively. Furthermore, negative logit values can distort the denominator when summed with other, positive logit values.

[17] https://github.com/shap/shap

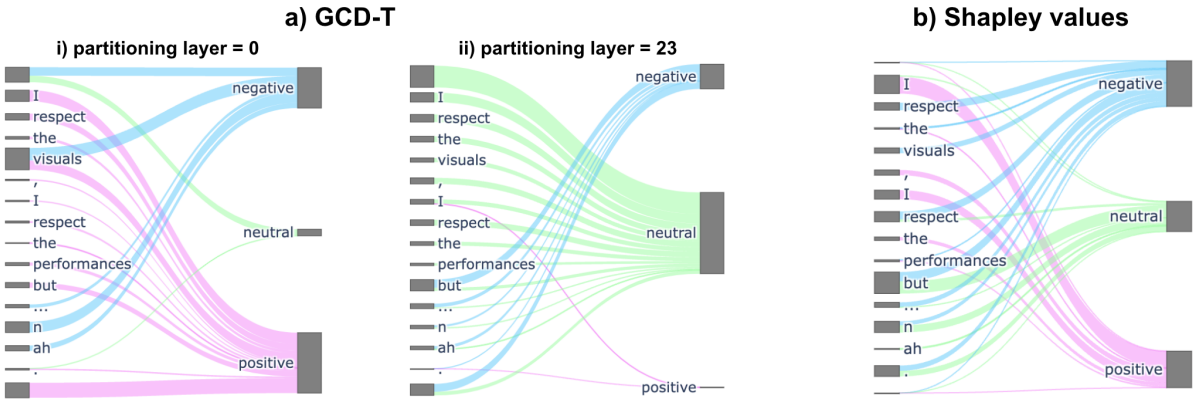


Figure 8: Example sentiment detection net-positive contributions to negative, neutral, and positive classes using a) GCD-T (layer-level decomposition) and b) Shapley values (model-level decomposition). The first and last token are the special tokens <*s*> and </*s*>.

layer-level decompositions, meaning `shap` gives less accurate estimations for large models like Transformers.

Figure 8 shows an example using a RoBERTa sentiment detection model (Loureiro et al., 2022) with a 1.5/5-star review for the movie *Joker: Folie à Deux* from the review aggregator website Rotten Tomatoes. With GCD-T, when components are partitioned before the first encoder layer (Figure 8a(i)), the start of the review ("I respect the visuals, I respect the performances") is interpreted as contributing a positive sentiment. When partitioning is done just prior to the final layer (Figure 8a(ii)), the words instead contribute neutrally to the sentiment. This is a result of the tokens getting contextualized across layers, likely because the review conclusion ("but... nah") forces a reinterpretation of the rest of the sentence. Conversely, the model-level Shapley value decomposition (Figure 8b) incorrectly shows "respect the visuals" and "respect performances" as contributing a negative sentiment.

Note that the model was originally trained to classify the sentiment of sequences, not individual tokens. Yet GCD-T still allows us to study the token level representational contributions learned by the model.[18] Because GCD-T does not require further model or probe training it is thus highly computationally efficient compared to other probing techniques. For instance, unlike Integrated Gradients (Sundararajan et al., 2017; Sanyal and Ren, 2021), GCD-T does not require the costly computation of gradients during the forward pass.

D SPRL Comparison to Prior Models

We compare the performance of RoBERTa and GPT-2 on the Decomp V1 dataset with models proposed by prior works: Logistic Regression (LR) (Reisinger et al., 2015), CRF (Teichert et al., 2017), and BiLSTM (Reisinger et al., 2015). Results are shown in Table 6. As we use the results reported by the original authors, the LR, CRF, and BiLSTM F1 scores are only for the sentences present in the original V1 dataset. Note that RoBERTa and GPT-2 models are additionally trained on Decomp V2 and passive sentences from our data augmentation process (described in §A), making the task more complex and the models more robust; however, there is a slight trade-off in performance on the original Decomp V1 dataset.

E Human Experiment Examples

For our human experiment, Table 5 shows an example set and Figure 9 shows the task given for the INANIMATE-ARG/Object arrangement.

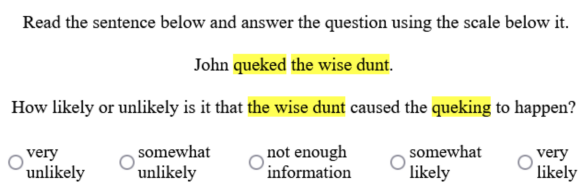


Figure 9: Example of display shown to participants.

[18] The original final classifier logits are equivalent to the summation of all GCD-T token contributions, plus the contribution of bias parameters.

Sentence	Condition	Position
John *wanted* the veect dunt.	PRED	Object
John queked the *wise* dunt.	MOD	
John queked the veect *bird*.	ANIMATE-ARG	
John queked the veect *book*.	INANIMATE-ARG	
The veect dunt *wanted* John.	PRED	Subject
The *wise* dunt queked John.	MOD	
The veect *bird* queked John.	ANIMATE-ARG	
The veect *book* queked John.	INANIMATE-ARG	

Table 5: Example set of stimulus in all arrangements. Emphasis not shown to participants.

	LR	CRF	BiLSTM	RoBERTa		GPT-2	
				Active	Passive	Active	Passive
instigation	76.7	85.6	88.6	84.4	90.6	82.8	92.2
volition	69.8	86.4	88.1	86.5	93.1	86.2	92.5
awareness	68.8	87.3	89.9	89.8	94.2	87.8	94.0
sentient	42.0	85.6	90.6	84.1	93.5	84.5	93.9
existed_before	79.5	84.8	85.1	86.9	89.4	85.9	86.8
existed_during	93.1	95.1	95.0	93.5	96.9	94.0	96.7
existed_after	82.3	87.5	85.9	87.1	90.3	85.9	89.0
Δ_state	54.6	66.1	71.0	67.5	69.6	67.0	70.1
Δ_possession	0.0	38.8	58.0	61.8	71.8	61.8	71.8
Δ_location	6.6	35.6	45.7	53.4	51.3	53.2	58.5

Table 6: Binary F1 scores of models on Decomp V1.