# Plausibility Processing in Transformer Language Models: Focusing on the Role of Attention Heads in GPT

**Soo Hyun Ryu**
Department of Psychology
University of Michigan
soohyunr@umich.edu

## Abstract

The goal of this paper is to explore how Transformer language models process semantic knowledge, especially regarding the plausibility of noun-verb relations. First, I demonstrate GPT2 exhibits a higher degree of similarity with humans in plausibility processing compared to other Transformer language models. Next, I delve into how knowledge of plausibility is contained within attention heads of GPT2 and how these heads causally contribute to GPT2's plausibility processing ability. Through several experiments, it was found that: i) GPT2 has a number of attention heads that detect plausible noun-verb relationships; ii) these heads collectively contribute to the Transformer's ability to process plausibility, albeit to varying degrees; and iii) attention heads' individual performance in detecting plausibility does not necessarily correlate with how much they contribute to GPT2's plausibility processing ability. Codes are available at github.com/soohyunryu/plausibility-processing-transformers

## 1 Introduction

Transformers are attention-based neural network models (Vaswani et al., 2017), which have brought breakthroughs in the field of Natural Language Processing achieving state-of-the-art performance in diverse downstream tasks. Such great performance is thought to be attributed to Transformers' ability to build dependencies even between long-distant words which attention heads are developed for (Merkx and Frank, 2021). To be specific, unlike previous neural network language models (e.g., Simple Neural Networks or Recurrent Neural Networks) that have issues retaining linguistic information coming from distant tokens, attention heads in Transformers enable to represent the meaning of tokens by integrating their contextual information without losing information from distant tokens (Bahdanau et al., 2015).

Provided that Transformer language models consist of multiple attention heads that serve different roles, previous studies examined functions that individual attention heads serve and how language processing work is divided inside Transformers (Clark et al., 2019; Voita et al., 2019; Vig, 2019; Jo and Myaeng, 2020). However, previous studies mostly focused on finding attention heads that process linguistic knowledge intrinsic to language systems such as morphosyntactic rules, and little attention has been paid to semantic knowledge, which requires much of world knowledge going beyond rules in language systems.

Consequently, we only have limited knowledge of how attention heads contribute to Transformers' general ability to process semantic knowledge. A number of studies (Bhatia et al., 2019; Bhatia and Richie, 2022; Ettinger, 2020; Han et al., 2022; Misra et al., 2020, 2021; Pedinotti et al., 2021; Peng et al., 2022; Ralethe and Buys, 2022) examined how Transformers process semantic knowledge in comparison with humans, but their focus was mostly on the models' performance from the final hidden state without answering where the specific type of knowledge is preserved or processed in Transformer models. A few studies started investigating how world knowledge is stored in Transformers (e.g., Meng et al. (2022) examined how GPT stores factual associations). However, the previous findings are yet generalizable to a broad range of semantic knowledge, and thus more studies are needed to understand how Transformers process other types of semantic knowledge.

In this regard, the present study aims to advance our knowledge of semantic knowledge processing in Transformer language models by closely investigating individual attention heads' ability in processing semantic plausibility and their causal contribution to Transformer's performance in plausibility processing. Among various types of plausibility, the especial focus of this paper is on the plausible



*Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 356–369
December 6-10, 2023 

relation between nouns and verbs. While recognizing the importance of considering a broader array of semantic knowledge in future studies, I made this specific choice because the objectives of the present paper are to demonstrate a set of attention heads can be specialized for specific type of semantic knowledge and to introduce a set of analyses that can be used to probe attention heads' role in processing semantic knowledge.

The semantic plausibility of the relationship between nouns and verbs can be determined by the degree to which semantic features of nouns and verbs match, as shown in sentences in (1) from Cunnings and Sturt (2018). For instance, in (1a), the syntactic dependent (*plate*) of the verb (*shattered*) has a feature [+shatterable], which builds a plausible relation with the verb (*shattered*). In (1b), however, the syntactic dependent *letter* does not have a feature [+shatterable], and thus it is semantically implausible dependent of the verb (*shattered*).

(1) a. Sue remembered the **plate** that the butler **shattered** ...
b. Sue remembered the **letter** that accidentally **shattered** ...

In order to examine how such knowledge is preserved and processed inside Transformer-based language models, this paper answers the following questions: (i) How similar are Transformer's plausibility processing patterns to humans'?; (ii) How sensitive is each of the attention heads in Transformers to plausibility relation?; and (iii) How do these heads make causal effects on Transformers' ability to process semantic plausibility?

After comparing patterns in plausibility processing between a group of Transformer-based language models and humans, it was found that GPT2 tends to process the plausibility between nouns and verbs in a way that is more similar to humans than other language model types. Several follow-up experiments that especially focus on GPT2 answered the last two questions. Specifically, it was uncovered that GPT2 has a set of attention heads that detect semantic plausibility, which are relatively diffusely distributed from the bottom layers to the top layers and that they exert causal effects on Transformers' semantic plausibility processing ability. GPT2's plausibility processing ability almost disappeared when the plausibility-processing attention heads are pruned, but the effects of removing a plausibility-processing attention head was not balanced nor proportional to the attention heads' performance in detecting plausible nouns. Rather, it was found that a single attention head accounts for most of plausibility processing ability of GPT2.

In what follows, I will provide a background that relates to the questions I address in this paper. In Section 3, I will compare Transformer-based language models' and humans' sensitivity to the plausibility of the relation between nouns and verbs. In Section 4, I will conduct an experiment to find attention heads that can detect semantic plausibility knowledge and examine how they are distributed inside the model. In Section 5, it will be examined how individual attention heads collectively make causal effects of on Transformers' sensitivity to plausibility. In Section 6, I will summarize the results and discuss the limitations of the study.

## 2 Background

**What roles do attention heads serve?** There have been a lot of studies that attempted to explain the language processing mechanism in Transformers with analyzing functions that distinct attention heads serve (Voita et al., 2019; Vig, 2019; Clark et al., 2019; Jo and Myaeng, 2020). Specifically, Voita et al. (2019) found attention heads specialized for a position, syntactic relation, rare words detection; Vig (2019) found attention heads specialized in part-of-speech and syntactic dependency; Clark et al. (2019) found attention heads specialized in coreference resolution; and Jo and Myaeng (2020) examined how linguistic properties at the sentence level (e.g., length of sentence, depth of syntactic trees and etc.) are processed in attention heads.

Despite numerous attempts in examining the roles of attention heads, the focus has been mostly on linguistic knowledge intrinsic to language systems which does not require much world knowledge that is indispensable for semantic knowledge processing. Thus, it needs to be closely examined how Transformers preserve and process such knowledge that facilitates sentence processing.

**How do we learn attention heads are specialized for certain linguistic knowledge?** In previous studies, attention heads are considered to be able to process a certain type of linguistic knowledge if attention distribution patterns in the attention heads are consistent with the linguistic knowledge (Voita et al., 2019; Vig and Belinkov, 2019; Ryu and Lewis, 2021). However, such regional analysis does not explain how much contribution attention

heads make to Transformers' ability to process linguistic knowledge because such information from the attention heads may fade away or be lumped along with the information flows - from bottom layers to top layers - eventually making little contribution to Transformers' ability to process the linguistic knowledge. Thus, to rigorously confirm the role of attention heads in processing a certain type of knowledge, it is crucial to analyze the causal effects that they make on Transformer's ability to process linguistic information (Belinkov and Glass, 2019; Meng et al., 2022; Vig et al., 2020).

In this sense, this paper will not only examine which attention heads can form attention distributions that are consistent with semantic plausibility knowledge, but also examine how much influence the attention heads can exert on Transformers' general ability to process plausibility.

## 3 Comparison between humans and Transformer language models in plausibility processing patterns

This section examines how a set of Transformer language models process plausibility of noun-verb relations in comparison with human data.

### 3.1 Data

In Cunnings and Sturt (2018), it was investigated how the degree of noun-verb plausibility affects the way humans process sentences. There are 32 sets of sentences with varying not only the plausibility of dependent-verb relations but also the plausibility distractor-verb relations[1].

(2)

a. *plausible - plausible*
... that the **plate** that the butler with the cup accidentally **shattered** ...

b. *plausible - implausible*
... that the **plate** that the butler with the tie accidentally **shattered** ...

c. *implausible - plausible*
... that the **letter** that the butler with the cup accidentally **shattered** ...

d. *implausible - implausible*
... that the **letter** that the butler with the tie accidentally **shattered** ...

[1]In experiments with language models, I removed sets of sentences whose tokens of interest are not recognized as a single token by the tokenizer.

### 3.2 Method

Cunnings and Sturt (2018) measured the degree of difficulty that people have when processing a certain noun-verb pair with reading times that are measured at verb[2] (*shattered* in (2)). To compare humans' responses with Transformer language models, I computed surprisals (Hale, 2001; Levy, 2008), also measured at verbs, as a metric that represents processing difficulty of the model, given a large set of evidence manifesting that surprisals computed from neural network language models can simulate human sentence processing patterns (Futrell et al., 2019; Michaelov and Bergen, 2020; Van Schijndel and Linzen, 2021; Wilcox et al., 2020).

Surprisal is a term that estimates the degree of the unexpectedness of tokens given their preceding context, which is computed by taking the negative log probability of a token conditioned on its preceding words (See Equation (A)). In neural network language models, the surprisal of a word is computed using the softmax-activated hidden state before consuming the word (Wilcox et al., 2018).

$$Surprisal(w) = -\log_2 P(w|h) \qquad \text{(A)}$$

where $h$ is the softmax-activated hidden state of the sentence before encountering the current word.

Both reading times and surprisals measured at verbs are expected to be greater in sentences with implausible nouns than in ones with plausible nouns since it is less likely to anticipate a certain verb after encountering a noun in an implausible relationship with the verb.

A set of Transformer language models to be tested includes ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), BERT (Kenton and Toutanova, 2019), and GPT2 (Radford et al., 2019). The versions of models that are tested have 144 attention heads, which are spread across 12 layers with 12 attention heads each. Models are accessed through Huggingface (Wolf et al., 2019).

### 3.3 Results

As shown in Figure 1, GPT2 exhibits the highest level of similarity to humans in processing the plausibility of noun-verb pairs, in comparison to other Transformer-based language models.

In addition, further statistical analysis using regression models supports GPT2's similarity with

[2]The original paper also talks about the spillover region following the verbs of interest, but this study focuses on the reading times (total viewing times) measured at the verb region.

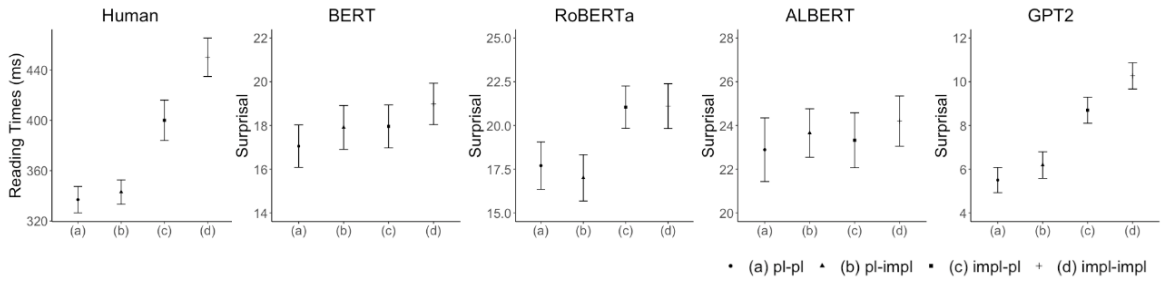



Figure 1: Surprisals computed from Transformer language models and reaction times from human subjects for processing different types of noun-verb pairs. Human reading times are from Cunnings and Sturt (2018). Shapes at the center and intervals for each condition represent means and standard errors.

humans in plausibility processing. First, significantly lower processing difficulties are observed when syntactic dependents are in a plausible relationship with the verb than when they are in an implausible relation for both human (estimate = .11 , SE = .01, $t$ = 9.26, $p$ < .001) and GPT2 (estimate = 4.81 , SE = .84, $t$ = 4.86, $p$ <.001).

Also, GPT2 showed marginally significant plausibility effects even with distractors that do not form a dependency relation with the verb (estimate = 1.57, SE = .84, $t$ = 1.87, $p$ = .06) (i.e., processing difficulties are greater in (b) and (d) than in (a) and (c)), similar to the human data where significant plausibility effects from distractors are found (estimate = .04, SE = .13, $t$ = 2.85, $p$ < .05)[3].

Being inconsistent with the human reading time data that show the interaction effects of dependent-plausibility and distractor-plausibility (estimate = .02, SE = .01, $t$ = 2.29, $p$ < .05), GPT2 data do not show significant interaction effects (estimate = .89, SE = 1.19, $t$ = .75, $p$ = .46). This absence of evidence for interaction effects in GPT2 may be due to the difference in sample sizes, which can impact the level of statistical significance. It would be possible to observe the interaction effects with the increased data size especially given a trend of interaction in GPT2: the surprisal difference between (a) and (b) is smaller than the surprisal difference between (c) and (d), consistent with human data. For the statistical results from other Transformer-based language models, see Appendix A.

[3]Plausibility effects observed for distractors in GPT2 and humans are due to the illusion of plausibility (Cunnings and Sturt, 2018): even distractors that cannot build syntactic dependency with cues (verbs) can be illusorily considered as the syntactic dependents, causing moderate plausibility effects while sentence processing.

### 3.4 Discussion

Compared to other language models, GPT2 is found to process plausibility between nouns and verbs in a similar way as humans do. While more rigorous study is required to explain the origin of GPT2's supeiror performance in simulating human plausibility processing patterns, I assume that the GPT2's similarity to humans arises from the psychological plausibility of its decoder-only architecture. In particular, it processes sentences incrementally much like the way humans process sentences (i.e., it constructs the meaning of a certain word only given its prefix, without any influence from the 'unseen' next coming words), unlike other types of language models that are tested exploit bidirectional processing (i.e., it process each word of sentences not incrementally, but integrating both preceding and following words.)

Given that GPT2 shows the most similar patterns as humans in processing plausibility of noun-verb relations, the following sections will examine the role that attention heads in plausibility processing, focusing on the GPT2 model.

## 4 Plausibility processing attention heads in GPT2

This section will examine whether GPT2 has a specific set of attention heads that can sensitively detect plausiblity of noun-verb relations, irrespective of syntactic dependency relation. Experimental stimuli were the same as previous experiment.

### 4.1 Method

In GPT2's attention heads, each token allocates different amounts of attention to previous tokens depending on the relevance of the two tokens[4].

[4]The relevance can be defined in terms of functions that attention heads serve. For instance, if an attention head is

With such a property of Transformers, the capacity of attention heads in detecting plausibility is measured in terms of *accuracy* that indicates how likely the plausible noun is to get higher attention than the implausible noun in a certain attention head (See Equation (B)).

$$Accuracy_{lh} = \frac{\sum_{j=1}^{k}[Attn(pl_j, v_j) > Attn(impl_j, v_j)]}{k} \quad \text{(B)}$$

, where *lh* refers to the location of attention heads (*h* for the *h*th head in the *l*th layer), *j* refers to the sentence id, $pl_j$ and $impl_j$ refer to the plausible and implausible nouns to be compared in the *j*th sentence set, $v_j$ refers to the verb in the *j*th sentence, and *k* is the number of sentence sets.

In order to ensure that the heads do not particularly work for tokens that form syntactic dependency but work for semantically related tokens, I measured the accuracy not only using pairs of syntactic dependents (*plate* vs. *letter* in (2)), but using pairs of distractors (*cup* vs. *tie* in (2)). Considering both of noun types enabled to find attention heads that can judge the plausibility between nouns and verbs regardless of syntactic compatibility between them. Thus, there are four comparisons between *plausible* and *implausible* conditions for each set of sentences: (pl-pl vs. pl-impl), (impl-pl vs. impl-impl), (pl-pl vs. impl-pl), (pl-impl vs. impl-impl), where the first and the second corresponds to syntactic dependents and distractors, respectively.

### 4.2 Results

I consider attention heads are able to process plausible relationships between nouns and verbs when their accuracy in identifying appropriate nouns surpasses the chance level, having the cutoff as 70% at my discretion. To select attention heads that can process the semantic plausibility regardless of the syntactic dependency relation between the noun and the verb, I consider attention heads whose accuracies are greater than 70% in both noun types.

With such criteria, eighteen attention heads are recognized to be able to process plausibility: [(0, 1), (0, 5), (0, 10), (1, 5), (1, 6), (1, 11), (3, 0), (4, 3), (4, 4), (4, 10), (5, 10), (5, 11), (6, 6), (7, 1), (7, 9), (8, 3), (8, 10), (9, 4), (10, 7)], where the first numbers refer to indexes of layers and the second refer to indexes of heads (i.e., (*i*, *j*) refers to the *j*th head in the *i*th layer.)) Among the attention heads that are found to process semantic plausibility, two attention heads - (1, 6) and (5, 10) - especially show noteworthy performance in detecting plausible, achieving 95% of accuracy. Please refer to Appendix B to see the values from each head.

### 4.3 Discussion

This section showed that a set of attention heads are particularly good at processing semantic plausibility between nouns and verbs. Such plausibility processing ability seems independent of their ability to process syntactic dependencies since their ability to process plausibility is not limited to processing syntactic dependents of verbs, but it is also applicable to distractors that do not form any syntactic dependencies with verbs.

Unlike attention heads specialized for processing a certain syntactic relation and superficial linguistic information such as word position or word rarity is clustered in a relatively small region (Voita et al., 2019), it seems that the components that process semantic plausibility are relatively evenly distributed across twelve layers and take up an even greater region: 18 attention heads out of 144 attention heads in the GPT2-small model. In the next section, it will be discussed how these plausibility-processing attention heads collectively exert causal effects on GPT2's plausibility-processing ability.

## 5 Causal effects of plausibility-processing attention heads on GPT2's plausibility sensitivity

In the previous experiment, attention heads capable of detecting plausible relations between nouns and verbs were found. The present section examines how such attention heads make causal influence on GPT2's sensitivity to plausibility between nouns and verbs. In particular, I attempt to answer two questions: (i) How GPT2's responses to plausible/implausible verb-noun pairs change when plausibility-processing attention heads are removed? and (ii) How does GPT2's plausibility-sensitivity change as attention heads are gradually pruned?

### 5.1 Influence of a set of plausibility-processing heads to plausibility sensitivity

In this study, I examine how GPT2's responses to plausible and implausible noun-verb relations

specialized for detecting *subject-verb* dependency relation, the amount of attention can reflect how likely two tokens are in the *subject-verb* relationship (Voita et al., 2019)

change when the plausibility-processing heads are removed.

#### 5.1.1 Method

Surprisals are computed from two models: i) GPT2 without plausibility-processing heads and ii) GPT2 after removing the same number of attention heads as i), but the heads to prune selected randomly. I included the random-removal model to see whether the disappearance of the plausibility sensitivity in GPT2 is simply attributed to taking away some portion of the information in GPT2, or it is caused by specifically removing plausibility processors. In order for reliability, we used 100 different random attention head sets for ii), and computed the average of surprisals from the 100 models.

Attention heads were pruned by replacing attention values with zeros, following Michel et al. (2019).

#### 5.1.2 Results

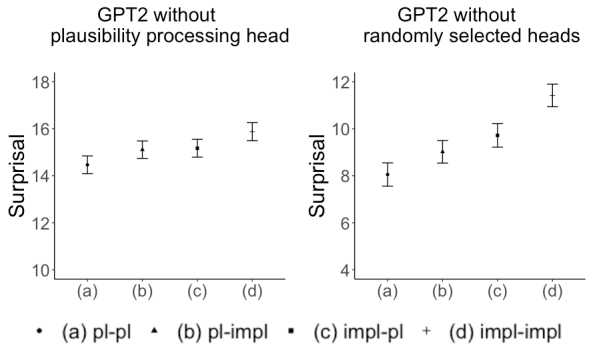



Figure 2: Surprisals computed from GPT2s after removing different sets of attention heads and reaction times from human subjects for processing different types of noun-verb pairs.

When removing the plausibility processing attention heads (left in Figure 2), no plausibility effects are found for syntactic dependents (estimate = .77, SE = .53, $t$ = 1.43, $p$ = .15) and for distractors (estimate = .71, SE = .54, $t$ = 1.32, $p$ = .19). Also, no interaction effects are found (estimate = 0.06, SE = 0.76, $t$ = 0.08, $p$ = 0.94)

Importantly, such a decrease is not the effect that is caused by simply removing some random components in GPT2. When randomly selected eight-teen attention heads are pruned (right in Figure 2), the GPT2 model better simulates human responses in processing plausibility. In this case, the significant plausibility effects are observed both in syntactic dependents (estimate = 2.40, SE = .69, $t$ = 3.46, $p$ < .001 ) and in distractors (estimate = 1.70, SE = .69, $t$ = 2.45, $p$ < .05), although interaction effects are not found as well (estimate = 0.73, SE = 0.98, $t$ = 0.75 $p$ = 0.46).

### 5.2 Gradual changes in GPT2's plausibility sensitivity as attention heads are pruned

The previous section examined how the set of plausibility-processing attention heads influences GPT2's responses to plausible or implausible noun-verb relations. Though it was shown that plausibility processing attention heads collectively contribute to GPT2's ability to process plausibility unlike other sets of attention heads, it is unanswered how individual attention heads contribute to GPT2's plausibility-processing ability. Do they have balanced contributions to GPT2's ability to process plausibility? Or, is it that only a small set of plausibility-processing attention heads account for most of the plausibility-processing ability of GPT2? In order to answer these questions, the following experiment investigates how GPT2's general sensitivity to plausibility gradually changes as attention heads are pruned one by one.

#### 5.2.1 Method

This study operationalizes GPT2's plausibility sensitivity as the difference in *surprisals* measured at the verbs of interest ('*shattered*' in (2)) in sentences with plausible nouns and in ones with implausible nouns as shown in Equation (C).

$$Plausibility\ Sensitivity = surprisal_{impl}(verb) - surprisal_{pl}(verb) \quad \text{(C)}$$

, where $\text{surprisal}_{\text{pl}}$(verb) and $\text{surprisal}_{\text{impl}}$(verb) refer to surprisals measured at the verb in a sentence with a plausible noun and in a sentence with an implausible noun, respectively.

I computed two plausibility sensitivities: one that compares surprisals at verbs when having plausible syntactic dependents of verbs in sentences and having implausible syntactic dependents ({(c)+(d)} - {(a)+(b)}) and the other that compares surprisals when having plausible distractors of verbs and implausible distractors ({(b)+(d)} - {(a)+(c)}).

Both types of plausibility sensitivities are measured at each point after gradually removing a plausibility processing attention head one by one. Attention heads were pruned in decreasing order of their accuracies[5] in detecting plausible nouns over implausible nouns.

[5] I used the average values of accuracies for dependents and for distractors that were computed in Section 3.

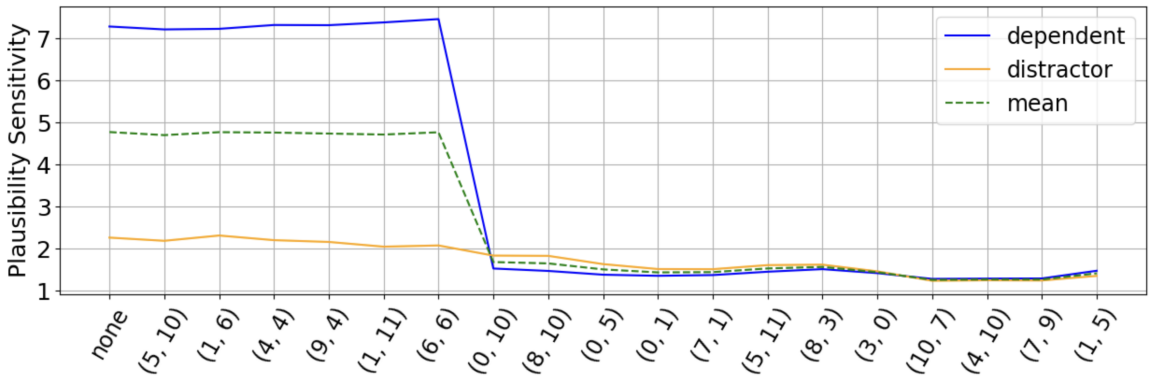



Figure 3: Changes in plausibility sensitivity by noun types as attention heads are gradually pruned. X-axis indicates plausibility-processing attention heads that are pruned at a certain point.

#### 5.2.2 Results

Figure 3 plots how the plausibility sensitivities for both types of noun-verb relations change as plausibility-processing attention heads are removed gradually.

When it comes to the plausibility sensitivity for distractors, the changes seem to be continuous. Such patterns suggest that the set of plausibility processing attention heads make a collective contribution to plausibility effects for distractors. Such collective contribution that plausibility processing attention heads make is especially supported by the fact that the gradual decrease in plausibility sensitivity over the course of removing 18 attention heads eventually led to the elimination of the statistically significant plausibility effects for distractors as observed in Section 5.1.

In contrast, the sensitivity to plausibility for the relation between syntactic dependents and verbs shows a drastic decrease upon the removal of the attention head (0, 10). The effect from the removal of the head (0, 10) shows that this particular head exerts a huge amount of causal effects on GPT2's general sensitivity to plausible relations between syntactic subjects and verbs[6]. Figure 4 confirms that the head (0, 10) causes a huge amount of causal contribution on GPT2's plausibility processing ability since it reduces the difference in surprisals between plausible conditions and implausible conditions, though it does not alone eliminate the significance in plausible effects for syntactic dependents (estimate = 1.29, SE = 0.61, $t$ = 2.10, $p < .05$) or for distractors (estimate = 1.40, SE = 0.61, $t$ = 2.29, $p < .05$).

One additional interesting finding is that the general level of surprisals upon the removal of the attention head (0, 10) increases considerably regardless of the condition. For instance, the removal of the single attention head (0, 10) increases surprisals by 2.79 bits on average across the four conditions, which seems to be huge given that the randomly selected 18 attention heads only led to the 1.89 bits of increase. Such trends indicate one possible explanation of the role of the head (0, 10): it contributes to GPT2's general ability to predict the next word, and such impact arises in any sentence, not only in the sentences that require plausibility-processing. In the next section, further analysis on the role of the attention head (0, 10) will be provided to address such a possibility.

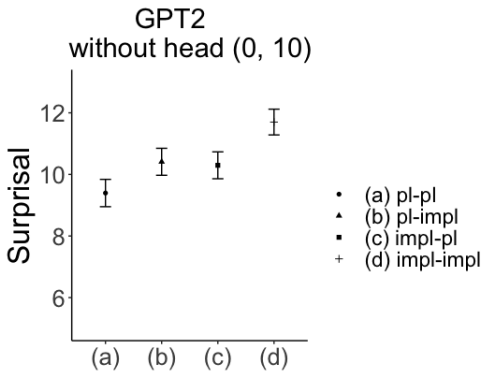



Figure 4: Surprisals by conditions computed with the GPT2 without a single attention head (0, 10)

### 5.3 Further analysis on the role of the attention head (0, 10)

To better understand the origin of GPT2's plausibility processing ability, the present study aims to further examine the role of (0, 10) that make great contribution to plausibility sensitivity in GPT2 . In

[6]The drastic drop after the removal of the head (0, 10) was also found when attention heads are removed in random order.

particular, I examine whether the (0, 10) is only specialized for semantic plausibility or is responsible for predicting next words in general sentences which leads to influence plausibility processing.

#### 5.3.1 Method

Perplexity in Equation (D) is the average value of surprisals computed from every tokens in corpus, which can be used to estimate the predictive power of language models in predicting next words given preceding context (Goodkind and Bicknell, 2018).

$$Perplexity(LM) = \frac{1}{M}\sum_{i=1}^{m} log_2 P(w_i|h) \quad \text{(D)}$$

, where $i$ is the index of words, $m$ is the number of words in corpus, and $h$ refers to the softmax-activated hidden state of the preceding context.

To examine how the general predictive power gets affected by the removal of the head (0, 10) in comparison with the removal of other heads, I computed the perplexities of GPT2 after removing each of 144 attention heads and compared those values. Andersen (1855)'s "The Money Box" story which has 41 sentences was used to compute perplexities.

#### 5.3.2 Results

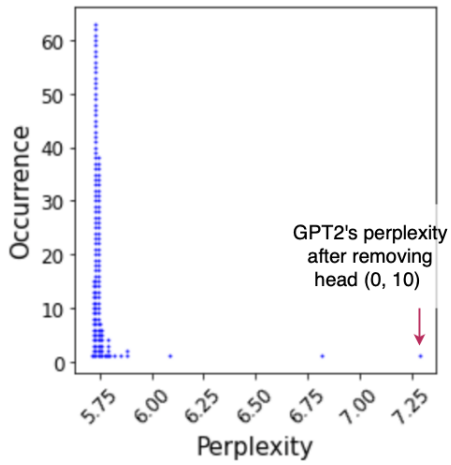



Figure 5: Histogram of 144 perplexities of GPT2, each of which is computed after removing single attention head

The perplexity of GPT2 with the entire set of attention heads was 5.47. In most of the cases, the removal of a single head does not seem to considerably affect GPT2's perplexity, since the perplexity remains to be in a similar range after the removal as shown in Figure 5[7]. However, it is clear that the removal of the head (0, 10) seriously harms the general predictive power of GPT2 because the perplexity becomes 7.27 after removing it, which is much greater compared to the most of other attention heads. This suggests that the head having the greatest influence on GPT2's plausibility processing ability is not specifically specialized for plausibility processing, but rather the attention head contributes to the general predictive power of any kind of sentence.

[7]For 95% of attention heads, the perplexities change by less than 0.1 bit after the removal.

### 5.4 Discussion

Results of this section suggest plausibility processing in GPT2 requires a collective contribution from a large set of plausibility processing attention heads, given that plausibility sensitivity decreases continuously as attention heads are gradually pruned.

At the same time, however, it was also shown that the amount of causal effects that each attention head makes are highly imbalanced since the attention head (0, 10), which contributes to GPT2's general predictive power, leads to a significantly more drastic decrease in plausibility sensitivity for dependents than other heads. Taken together, although a single attention head can account for a great portion of the plausibility effects, other plausibility-processing attention heads make an additional contribution to GPT2's plausibility-processing ability.

Interestingly, the head (0, 10) did not achieve noteworthy performance in detecting plausible nouns over implausible nouns in Section 4. This suggests that analyzing the causal effects each attention head makes is essential to understanding the role that attention heads serve, provided that the performance that each attention head shows in processing particular linguistic information does not necessarily align with how much it contributes to the model's performance in processing the specific information.

In addition, how the plausibility-processing attention heads affect Transformers' general ability needs to be investigated in relation to other attention heads that are specialized for different linguistic knowledge. This is especially the case given the findings that the way plausibility sensitivity decreases along with the gradual heads-pruning varies by the relation types that nouns build with verbs (i.e., syntactic dependents or distractors), which must be handled by different attention heads.

## 6 Conclusion & Limitations

The present study has shown how semantic plausibility is processed in Transformer language models, especially focusing on the role of attention heads. First, I demonstrated that GPT2, whose decoder-only architecture is more aligned with the way humans process sentences, shows greater similarity to humans in plausibility processing compared to other Transformer-based models such as BERT, RoBERTa and ALBERT. Then, a series of experiments showed a set of attention heads are found to process plausibility, and those heads are diffusely distributed across 12 layers in GPT2. Moreover, it was observed that they make imbalanced but collective causal contributions to GPT2' plausibility-processing ability, which establishes the importance of causal effect analysis in attention-head-probing studies.

Although the results provide a window into how Transformers process semantic knowledge of plausibility, this study has a few limitations to be addressed in future studies. First, the scope of the study is restricted to the plausibility of noun-verb relations although there exist many different types of semantic knowledge. This limitation stems from the present paper's intention to 'initiate' an exploration of Transformers' attention heads in handling of semantic knowledge and to exploit diverse and robust techniques for the exploration, rather than serving as a definitive endpoint that accounts for an exhaustive set of semantic knowledge. However, future investigations should expand the current study's scope for better generalizability.

Moreover, the study does not detail how attention heads interact with other components like hidden states across layers or feed-forward perceptrons. Such details would be essential in enhancing our understanding of the attention head roles in plausibility processing by elucidating how these heads impact Transformer models' plausibility processing ability. As such, subsequent studies should delve deeper into these interactions for a more accurate understanding of their role in semantic knowledge processing.

As these limitations are addressed, I anticipate further advancements in explaining Transformer models' capacity for semantic knowledge processing, founded on the novel findings and methods introduced in this study.

## Acknowledgements

This research took place as part of EECS 595 Natural Language Processing, a course taught by Joyce Chai at the University of Michigan in the fall term of 2022. I am truly grateful for the invaluable insights shared by all of my class instructors. Additionally, I extend my gratitude to the members of the Computational Cognitive Science Lab at the University of Michigan - Richard Lewis, Logan Walls, Yuxin Liu, Andrew McInnerney, Sean Anderson and Sarah Liberatore - for their instructive suggestions and guidance. I am also deeply appreciative of the four anonymous reviewers at the ACL Rolling Review for their productive feedback, which significantly enhanced the quality of the paper.

## References

Hans Andersen. 1855. *Hans Andersen's Fairy Tales: The money box*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Sudeep Bhatia and Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Sudeep Bhatia, Russell Richie, and Wanling Zou. 2019. Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29:31–36.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Ian Cunnings and Patrick Sturt. 2018. Retrieval interference and semantic interpretation. *Journal of Memory and Language*, 102:16–27.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019.

Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Simon Jerome Han, Keith Ransom, Andrew Perfors, and Charles Kemp. 2022. Human-like property induction is a challenge for large language models.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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 preprint arXiv:1907.11692*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*.

Danny Merkx and Stefan L Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2021. Do language models learn typicality judgments from text? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 1–11.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035.

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

Sello Ralethe and Jan Buys. 2022. Generic overgeneralization in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196.

Soo Hyun Ryu and Richard L Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Table 1: Statistical analysis on plausibility effects in human and Transformer-based language models.

| | | Human | BERT | RoBERTa | ALBERT | GPT2 |
|---|---|---|---|---|---|---|
| Difficulty measurement | | reading times | surprisals | | | |
| Plausibility effects (syntactic dependents) | estimate | .11 | 1.10 | 4.11 | .55 | 4.81 |
| | SE | .01 | 1.38 | 1.83 | 1.77 | .84 |
| | *t* | 9.26 | .78 | 2.24 | .31 | 4.86 |
| | *p* | **<.001** | .44 | **<.05** | .76 | **<.001** |
| Plausibility effects (distractors) | estimate | .04 | 1.03 | .06 | .87 | 1.57 |
| | SE | .13 | 1.38 | 1.83 | 1.77 | .84 |
| | *t* | 2.85 | .75 | .03 | .49 | 1.87 |
| | *p* | **<.05** | .46 | .97 | .62 | **< .10** |
| Interaction effects ( dependents × distractors) | estimate | .02 | .17 | .76 | .11 | .89 |
| | SE | .01 | 1.95 | 2.59 | 2.50 | 1.19 |
| | *t* | 2.29 | .09 | .29 | .04 | .75 |
| | *p* | **<.05** | .93 | .77 | .96 | .46 |

## A Statistical analysis on plausibility effects

In order for quantitative analysis on how well Transformer language models simulate plausibility effects found in human data (Cunnings and Sturt, 2018), linear regression models for language model data were fit with the following equation: $surprisal \sim subject\_plausibility * distractor\_plausibility$.

The results are shown in Table 1. Results for human data are from Cunnings and Sturt (2018).

## B Scores for detecting the plausible noun-verb relations by attention heads

The performance of attention heads in selecting the plausible nouns in relation with verbs over the implausible ones was measured in terms of *accuracy* in the main text. The details of the method are provided in Section 4.

In addition to accuracy, I also computed attention differences which indicate how much more attention values plausible nouns get compared to implausible nouns (See Equation (E)). The attention differences obtained from all attention heads are shown in Figure 6.

$$Attention\ Difference_{lh} = \sum_{j=1}^{k}[Attn(pl_j, v_j) - Attn(impl_j, v_j)] \quad \text{(E)}$$

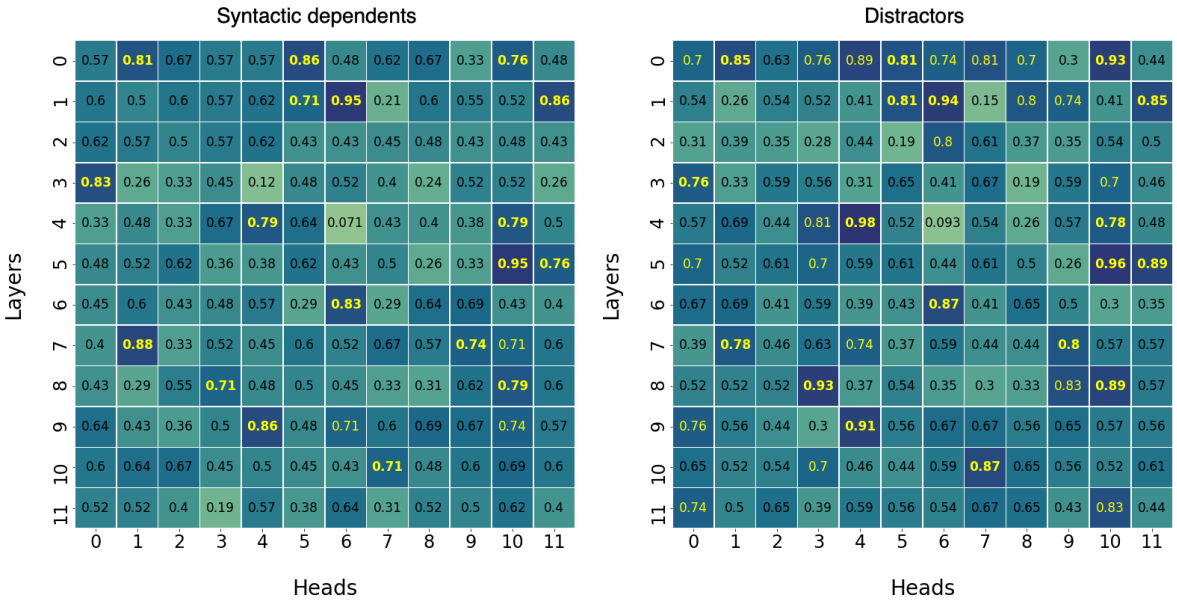
,where *lh* refers to the location of attention heads (hth head in the *l*th layer), *j* refers to the sentence id, $pl_j$ and $impl_j$ refer to the plausible and implausible nouns to be compared in the *j*th sentence set, $v_j$ refers to the verb in the *j*th sentence, and *k* is the number of sentence sets.

Metrics were computed two times: one by comparing plausible syntactic dependents and implausible syntactic dependents, and the other by comparing plausible distractors and implausible distractors.

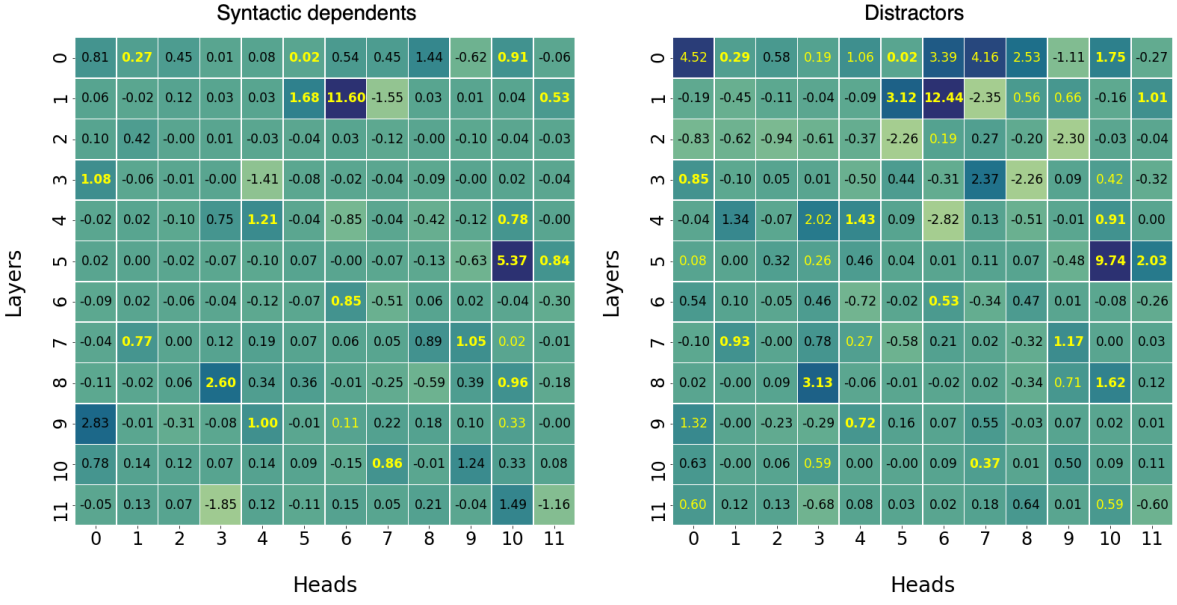
## C Changes in surprisal values as attention heads are gradually pruned

In Section 5.2, it was observed how the plausibility sensitivity changes as the plausibility-processing attention heads are gradually pruned. To provide additional information, this section shows how the surprisals for each condition change along with the gradual head-pruning process.

Surprisals were computed at the verb for each sentence in Cunnings and Sturt (2018)'s experimental data. The metrics were computed multiple times after removing one of the plausibility-processing attention heads. The computed surprisal values were then averaged by conditions. The plot that shows how surprisal values change by conditions is given in Figure 7.




(a) Accuracy




(b) Attention Difference

Figure 6: Accuracy and attention difference by attention heads. Attention heads annotated with bold-yellow are with accuracy greater than 0.70 in both subjects-comparison and distractors-comparison and thus considered to be specialized for plausibility processing; Attention heads annotated with non-bold-yellow are the ones that showed accuracy greater than 0.70 only for the corresponding condition; Attention heads annotated with black are found to be insensitive to plausibility (accuracies are less than 0.7 for both noun types).

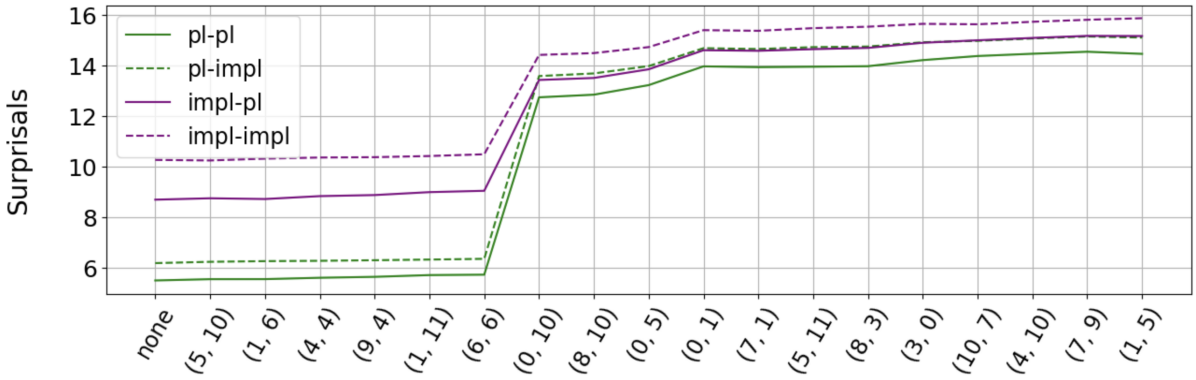



Figure 7: Changes in surprisals by conditions as attention heads are gradually pruned. X-axis indicates plausibility processing attention heads that are pruned at a certain point. Attention heads were removed in decreasing order of accuracies in selecting plausible nouns over implausible nouns.