Language Models as the Brain: Probing Linguistic Structures within Neural Language Models using Minimal Pairs

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, Jonathan R. Brennan
University of Michigan, ShanghaiTech University, LMU Munich
Ann Arbor, MI, USA, Shanghai, China, Munich, Germany
{linyangh, jobrenn}@umich.edu
{chenpl, liyn2}@shanghaitech.edu.cn
nie@cis.lmu.de

Abstract

Inspired by cognitive neuroscience studies, we introduce a novel 'decoding probing' method that uses minimal pairs benchmark (BLiMP) to probe internal linguistic characteristics in neural language models layer by layer. By treating the language model as the brain and its representations as 'neural activations', we decode grammaticality labels of minimal pairs from the intermediate layers' representations. This approach reveals: 1) Self-supervised language models capture abstract linguistic structures in intermediate layers that GloVe and RNN language models cannot learn. 2) Information about syntactic grammaticality is robustly captured through the first third layers of GPT-2 and also distributed in later layers. As sentence complexity increases, more layers are required for learning grammatical capabilities. 3) Morphological and semantics/syntax interface-related features are harder to capture than syntax. 4) For Transformer-based models, both embeddings and attentions capture grammatical features but show distinct patterns. Different attention heads exhibit similar tendencies toward various linguistic phenomena, but with varied contributions. Notably, specific heads consistently offer the most contributions.

Keywords: minimal pairs, probing, linguistic capabilities of neural language models

1. Introduction

Given the exceptional performance of self-supervised language models in a wide range of NLP tasks (Brown et al., 2020; Devlin et al., 2018; Radford et al., 2019), their ability to capture linguistic information has piqued the interest of many cognitive neuroscientists. These models offer new insights and methodologies for studying the neural mechanisms underlying language processing.

Notably, encoding analysis in cognitive neuroscience, which leverages linguistic representations from language models to reconstruct neural signals when corresponding stimulus presented, has gained popularity as a method to identify the neural substrates of speech and language processing in the brain (Schrimpf et al., 2021; Li et al., 2022; Goldstein et al., 2022). However, these studies have not clearly delineated the specific linguistic information that neural language model representations capture when encoded in the brain; conclusions thus far point mostly to the contextual aspects of language processing. Therefore, determining the exact linguistic content captured by these neural language model representations is of paramount importance.

Another line of research has delved into probing linguistic capabilities. Marvin and Linzen (2018) introduced the concept of targeted syntactic evaluation. They designed minimal sentence pairs where two sentences differ by only one word, yet this difference renders one sentence acceptable and the other not. Here's an example:

(1) *Simple agreement*:

a. The cats annoy Tim. (*grammatical*)
b. *The cats annoys Tim. (*ungrammatical*)

If a language model assigns a higher probability to the acceptable sentence over the unacceptable one, it is considered to have performed correctly on this task. Building on this approach, Warstadt et al. (2020) introduced a comprehensive minimal pairs benchmark called BLiMP covering a broad range of grammatical phenomena. Such an extensive benchmark offers valuable insights into the holistic linguistic comprehension of a language model and help draw more concrete conclusions about the specific kinds of grammatical features learned in a certain model.

However, previous studies using minimal pairs only focus on the probability assigned by the entire model (i.e., the final layer), not delving deeply into the linguistic characteristics inherent within the neural language model (i.e., the intermediate layers). Conversely, many encoding analysis studies in cognitive neuroscience favor the use of intermediate layer embeddings over solely the final layer. For instance, Caucheteux et al. (2021) indicates that the 2/3rd layer achieves the peak encoding correlation score, while Schrimpf et al. (2021) reveals variability in the optimal layer across participants. Given these insights, probing the linguistic information in intermediate embeddings of language models becomes crucial.

Several studies have explored the linguistic properties encapsulated in intermediate embeddings

from other perspectives. Tenney et al. (2019) examined the structural insights gleaned from contextual word representations, but their approach combined information from multiple layers rather than probing each layer individually. Hewitt and Manning (2019) and Manning et al. (2020) further identify the emergent syntactic structures within neural networks. However, a gap still remains in comprehensively probing specific linguistic nuances across model layers, an aspect our study aims to address using minimal pairs for a more granular analysis.

Given focus on intermediate layers, the traditional method of comparing probabilities assigned by the entire language model becomes infeasible. This is because the probabilistic output of LM is based on the final layer processed through the softmax function. Although one could apply a softmax to an intermediate layer, there are inherent challenges. These layers aim to capture linguistic nuances rather than provide direct vocabulary probability distributions. Additionally, since they aren't optimized during training for token prediction, introducing softmax might distort the true linguistic patterns they represent. As such, a new probing method for these layers is essential.

In cognitive neuroscience, when subjects are exposed to stimuli of different conditions, if stimulus categories or labels can be correctly categorized based on features derived from neural signals, it is inferred that those neural signals carry task-relevant representational information. (Haxby et al., 2001; Mitchell et al., 2008; Sudre et al., 2012). Drawing inspiration from this, we treat the neural language model as a "brain". By feeding minimal sentence pairs into the language model simultaneously, we can obtain the internal activations corresponding to whatever distinguishes these stimuli. We designed a simple yet effective decoding probing method using these 'activations' to decode grammatical or ungrammatical labels.

Decoding probing offers a precise lens to examine the linguistic intricacies within each layer of neural language models. By applying this decoding method along with the large minimal pairs benchmark, we can pinpoint which layers excel at capturing morphology, syntax, semantics, or other linguistic phenomena. This granularity provides insights into how the model processes language hierarchically. In essence, decoding probing unlocks a deeper understanding of the inner workings of neural language models.

In our present study, we employ decoding probing to investigate GPT-2, ELMo, and GloVe. These models exemplify self-supervised, RNN-based, and word embedding language models, respectively. Our exploration yields four principle results: 1) We first confirm that these intermediate layers support grammatical decoding that is not possible with simpler RNNs and GLoVE. 2) GPT-2 XL learns syntactic grammaticality through its initial layers, with this information being encoded among later layers. As sentences grow in complexity, more layers are needed to capture grammatical information. 3) Morphology and semantics-related features are captured by later layers than purely syntactic ones in LMs. 4) For Transformer-based models, while both embeddings and attentions capture grammatical attributes, they exhibit different patterns. Intriguingly, while different attention heads perform similarly toward various linguistic phenomena, a limited set of heads consistently emerge as predominant contributors.

2. Decoding Probing

2.1. Decoding analysis in the brain

Decoding analysis has emerged as a cornerstone methodology in cognitive neuroscience, enabling researchers to infer the specific representational content encoded in neural patterns (Kriegeskorte et al., 2006). This approach primarily revolves around the concept of training classifiers to predict specific stimulus conditions (e.g., a particular image or word) based solely on neural responses. When a classifier can predict a stimulus condition with accuracy significantly above chance, it suggests that the neural data contains information specific to that condition (Norman et al., 2006). The foundational idea posits that if specific stimulus attributes can be consistently decoded from neural configurations, such configurations must inherently represent those attributes (Haynes and Rees, 2006).

One of the paradigmatic works leveraging decoding analysis in language is Wehbe et al. (2014). In this study, the authors employed fMRI to measure neural responses in subjects as they read nouns. Through the lens of decoding, the researchers could distinguish between neural patterns elicited by different nouns, underscoring the perceptual and semantic distinctions between them.

2.2. Decoding probing in LMs

Building on this foundational idea and to determine the linguistic competence of each layer within the neural language models, we propose a decoding probing approach (Figure 1). For every minimal pair, we extracted embeddings and attentions as LMs' 'activations', and we trained a binary classifier to decode whether a given sentence was grammatical or ungrammatical. This approach allowed us to investigate if specific internal regions of the model contains information related to the grammaticality of that particular minimal pair.

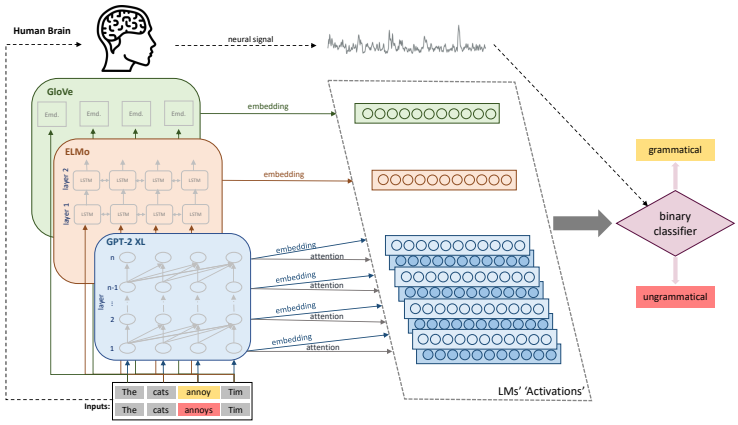

Figure 1: Decoding probing pipeline. Note: in the current study, we didn't include experiments on the brain; it's just for demonstrating the core idea as the dashed lines also imply.

3. Experimental Setup

3.1. Minimal pairs as model inputs

We use minimal pairs (sentences with subtle but grammatically significant differences) as the basis for grammatical assessment. By comparing the assigned probabilities to grammatical and ungrammatical utterances in these pairs, we can probe if the language models truly grasp linguistic rules or merely rely on statistical biases. This method, providing a precise perspective on the linguistic comprehension of LMs, has been adopted in prior research (Marvin and Linzen, 2018; Wilcox et al., 2018; Gauthier et al., 2020). We employed the Benchmark of Linguistic Minimal Pairs (BLiMP) dataset by Warstadt et al. (2020). BLiMP is the largest and most comprehensive such minimal pairs dataset, comprising 67 tasks covering 12 phenomena. These phenomena are divded into "morpholgoy", "semantics-syntax interface' and "syntax" categories. Each task presented 1, 000 sentence pairs. Within each pair one pair is grammatical and one is not. See Table 1 for details.

3.2. Model

GloVe acts as a symbolic static word embedding that encodes words into fixed vectors based on global word co-occurrence statistics regardless of specific context (Pennington et al., 2014).

ELMo is based on the LSTM architecture, which represents a type of dynamic, context-sensitive embeddings (Sarzynska-Wawer et al., 2021). For our analysis, given that ELMo's default word representation is a linear combination of two bi-LSTM layers and one char-level CNN embedding, and our intent is to investigate the performance of each individual layer, we manually extracted the LSTM activations instead of employing the standard embeddings.

GPT-2 XL is a self-supervised model (Radford et al., 2019). With its Transformer architecture, GPT-2 XL is known for its ability to generate coherent text and understand complex language structures. There are 1 token-level embedding layer and 48 contextual Transformer layers in the GPT-2 XL. For each Transformer layer, there are 20 attention heads. We used both hidden states and the attention matrices to do later decoding probing.

3.3. Internal Activations

Sentence Embedding When inputting sentences into various language models, we obtain word embeddings for each token in the sentence. The strategy for aggregating these embeddings into a coherent sentence representation varies depending on the model used. Given that GloVe provides static word embeddings, we employed a bag-of-words (BoW) model to construct the sentence representation. We computed the mean of the embeddings of all tokens within a sentence to generate a sentence representation. For ELMo and GPT-2 XL, the representation of the last token in each sentence was extracted from each layer to serve as the sentence representation for that particular layer. This decision is grounded in the understanding that the last word's embedding, especially in context-sensitive models like ELMo (a bidirectional model) and GPT-2 XL (a unidirectional model), encapsulates substantial contextual information from the entire sentence. The choice of the last word ensures that the representation has been influenced by all preceding tokens.

Attention We also investigate how attention matrices in GPT-2 XL captures grammatically information. For each attention head, we vectorized its attention matrix. To analyze the collective behavior,

we concatenated these vectorized matrices across all attention heads to produce a final attention vector. Additionally, we examined the behavior of individual attention heads by probing their respective vectorized matrices without concatenation.

3.4. Setup and Metrics

10-fold CV To ensure the robustness and generalizability of our results, we adopted a 10-fold cross-validation strategy to train the logistic regression classifier. The final evaluation metric was the averaged F1 score on the held-out set across all ten validation folds.

Feature Capture Depth We explored the number of layers GPT-2 XL requires to grasp the linguistic characteristics for a particular task. To mitigate potential noise and obtain a more robust measure, we defined this layer as the first instance where the F1 score reaches 99% of the maximum F1 score observed across all layers for that task.

Sentence Complexity We measured sentence complexity, using the depth of its syntactic tree. We normalized the syntax tree depth by the length of the sentence.

$$\text{Sentence Complexity} = \frac{\text{Depth of Syntax Tree}}{\text{Sentence Length}}$$

The depth of the syntax tree was computed using the SpaCy (Honnibal and Montani, 2017) package in Python. We then correlate sentence complexity with the depth at which GPT-2 XL effectively learns the underlying linguistic patterns.

4. Results

4.1. BoW and RNN-based sentence embeddings distinguish grammaticality for some, but not all, minimal pairs

For each model, we utilized the averaged F1 score across 10 cross-validation folds to denote each layer's performance. The highest score across all layers is the final F1 score for the model. We did a comprehensive probing analysis for GPT-2 XL, ELMo, and GloVe across all 67 linguistic tasks. A more focused result can be found in Figure 2, which shows average performance on 12 linguistic grammatical phenomena. The top four panels represent morphology probing results, the middle four showcase the semantics-syntax interface, and the bottom four are dedicated to syntax. This result shows that GPT-2 XL consistently stands out, delivering superior performance across all language categories. ELMo, while lagging behind GPT-2 XL in morphology and semantics/syntax interface, shows significant closeness to GPT-2 XL for the

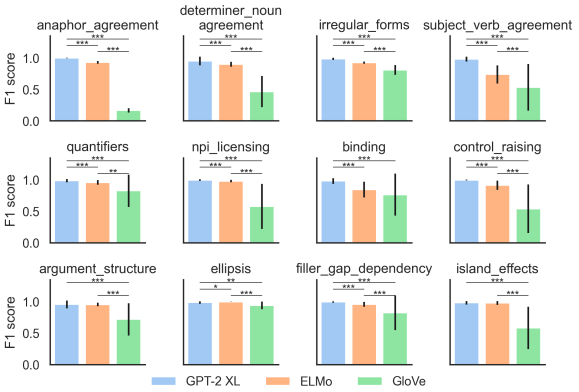

Figure 2: Decoding probing results of GPT-2 XL, ELMo and GloVe. The symbols '***', '**', and '*' denote t-test p-values less than 0.001, 0.01, and 0.05, respectively.

syntatic benchmarks. Interestingly, despite its rudimentary bag-of-words approach in sentence representation, GloVe performs well in specific areas such as quantifiers and ellipses. In some tasks, GloVe is even better than ELMo.

In the subsequent sections, we will narrow our focus to GPT-2 XL. We also excluded tasks (26 in total) where GloVe's F1 score is higher than 0.9, trying to delve deeper into the unique mechanisms by which self-supervised language models capture linguistic characteristics.

4.2. Feature capture depth in GPT-2 XL

4.2.1. Grammaticality information captured through the first third of GPT2-XL layers

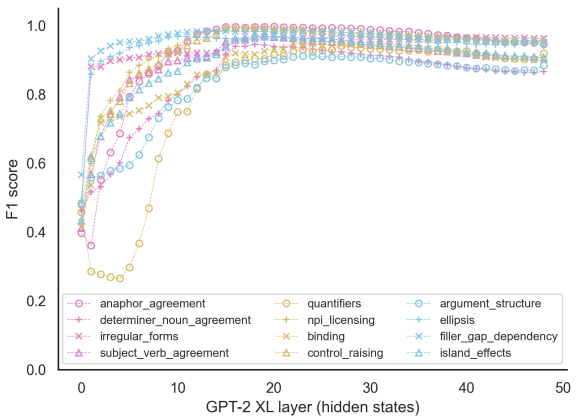

Figure 3: Decoding probing results of GPT-2 XL. Pink lines represent morphology, yellow lines highlight the semantics/syntax interface, and blue lines correspond to syntax.

Level	Phenomenon	N	Grammatical Example	Ungrammatical Example
Morphology	Anaphor Agreement	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Morphology	Determiner Noun Agreement	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Morphology	Irregular Forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Morphology	Subject Verb Agreement	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*
Semantics	Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Semantics/Syntax	NPI Licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Semantics/Syntax	Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Semantics/Syntax	Control Raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Syntax	Argument Structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Syntax	Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Syntax	Filler Gap Dependency	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Syntax	Island Effects	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*

Table 1: BLiMP dataset details. This table is adapted from Warstadt et al. (2020). N denotes how many tasks are within each linguistic phenomenon. For each task, there are 1,000 pairs of sentences.

After we filtered out tasks where GloVe's F1 score is higher than 0.9, we retained tasks where GPT-2 XL's self-supervised mechanism contributes uniquely. As shown in Figure 3, grammatical information appears to be predominantly captured gradually through the first third of the GPT-2 XL layers. These linguistic features are also captured by subsequent layers, but with a slight downward trend.

4.2.2. Complex sentences require more layers to capture linguistic information

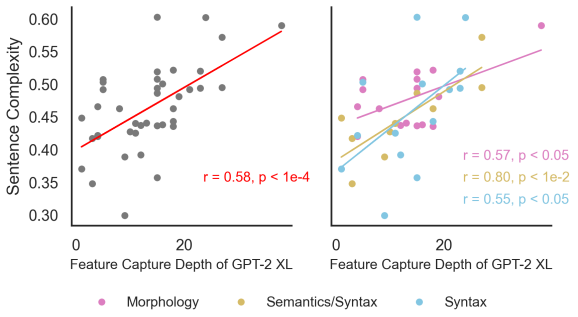

Figure 4: Sentence complexity and feature capture depth shows strong linear correlation in GPT-2 XL.

We also studied the interplay between sentence complexity and the depth required by the model to capture the corresponding linguistic information. Analyzing the all 41 tasks (after excluding 26 tasks where BoW sentence embedding model based on GloVe excelled), we compared sentence complexity against the feature capture depth for each task. The left panel of Figure 4 illustrates this relationship. A distinct linear relationship ($r = 0.58$, $p < 1 \times 10^{-4}$) emerges between sentence complexity and the required depth for feature capture. As the complexity of the sentences increases, GPT-2 XL demands more layers to effectively internalize the associated linguistic features.

4.3. Semantics-syntax interface and morphology are harder to learn than syntax for LMs

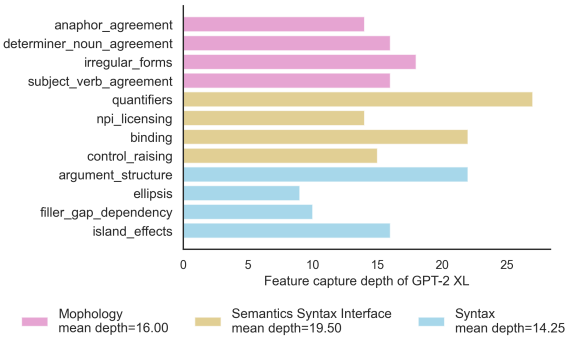

Figure 5: Feature capture depth for 12 linguistic phenomena, derived from 41 tasks after filtering out those where GloVe's F1 score exceeds 0.9.

Figure 5 illustrates the feature capture depth for various linguistic phenomena in GPT-2 XL. This depth is defined as the number of layers required for the model to achieve 99% of the maximum F1 score for each phenomenon, as described in the Methods section. Here, we observe that, on average, the semantics-syntax interface demands the highest depth for efficient feature capture, suggesting its inherent complexity. Morphology follows closely, while pure syntactic phenomena can be captured through a relatively shallower depth.

Figure 6 further illustrates this finding. Here, the average feature capture depth required to achieve various thresholds of the maximum F1 score is plotted for the three linguistic levels. The curve corresponding to the semantics-syntax interface again consistently sits at the top, indicating that it requires more layers to reach comparable performance thresholds. Morphology occupies the middle ground, while syntax consistently requires the least depth. Roughly speaking, for GPT-2 XL to capture a comprehensive understanding of

the semantics-syntax interface, it requires approximately 20 layers. For morphology, this number is around 16 layers, while syntax, being the most straightforward, needs roughly 14 layers.

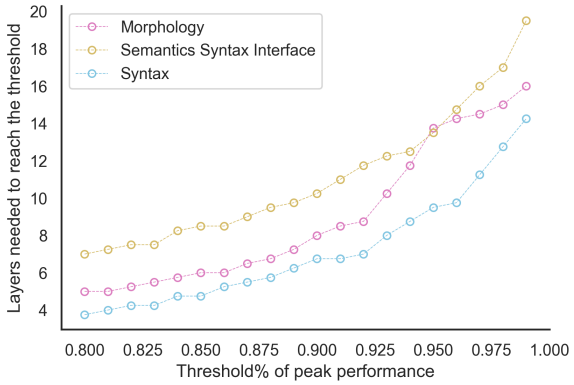

Figure 6: Average feature capture depth required by GPT-2 XL for different linguistic levels to reach certain threshold of the peak performance.

To supplement these findings, the right panel of Figure 4 presents linear fits between sentence complexity and feature capture depth for each of the three linguistic levels. A clear linear relationship is evident across all three, with the semantics-syntax interface exhibiting the most pronounced correlation. Intriguingly, the slope of the linear fit offers insights into the relative difficulty of capturing linguistic features for each category. A gentler slope indicates that more layers are necessitated per unit increase in sentence complexity. Both morphology and the semantics-syntax interface showcase flatter slopes compared to syntax, reaffirming that these two linguistic levels present steeper learning challenges for GPT-2 XL.

Results in 4.1 comparing GPT-2 XL, ELMo and GloVe also suggest that features related to morphology and the semantics/syntax interface are more challenging to capture compared to syntax.

4.4. Attention distribution across many heads supports grammatical representations, but not much on morphology

We examined the attention patterns in GPT-2 XL across its layers. Notably, attention mechanisms exhibited weaker performance in capturing morphology than in recognizing syntax or the semantics-syntax interface. Unlike the embeddings (or hidden states) which exhibit a clear gradual or incremental pattern in information capture, the concatenated attention matrices from all 20 heads did not display such a trend during decoding probing as shown in Figure 7. This suggests that

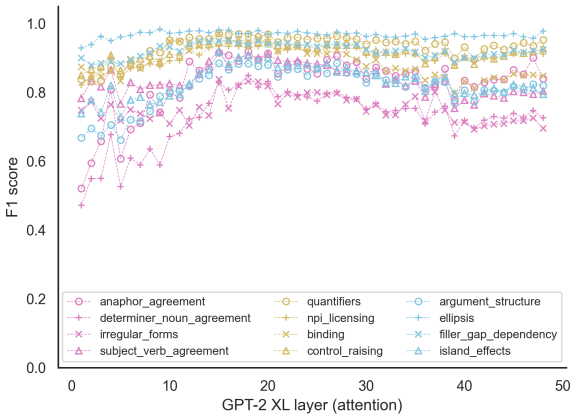

Figure 7: Decoding probing results of GPT-2 XL's attention matrices.

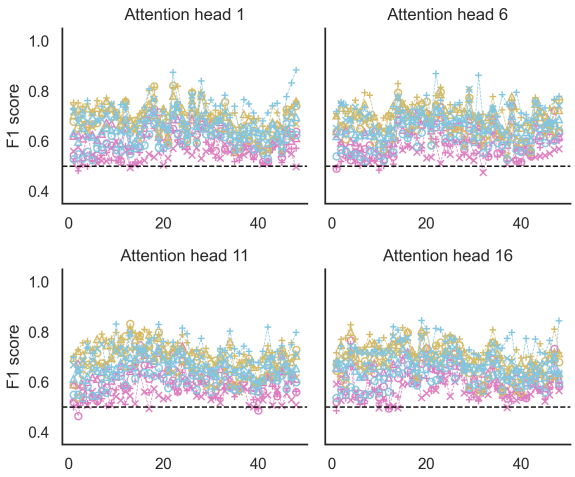

Figure 8: Decoding probing results for single attention heads. We take 4 heads as the examples. Legend is the same as in Figure 7.

the each attention layer might operate somewhat independently, different from hidden states building upon information from previous layers. In addition, individual attention head detection strengthens this observation. The layer-by-layer performance does not show an incremental pattern, but has significant oscillations, as shown in Figure 9.

In our exploration of individual attention heads in Figure 9, we observed that most scores falling below 0.8. However, when the attention vectors from all heads were concatenated and examined collectively, there was a notable surge in the F1 score, exceeding 0.9. This suggests that while individual attention heads may capture specific facets of syntactic information, a comprehensive representation emerges only when the information from all heads is integrated.

The left panel of Figure 9 shows that various

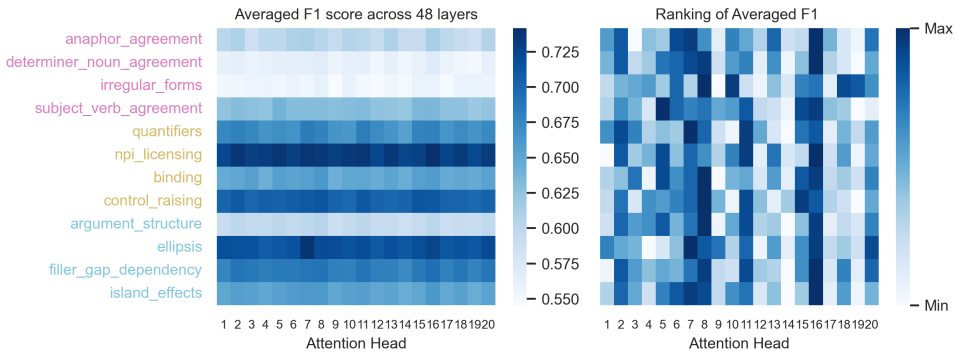

Figure 9: Performance (F1 score) and ranking of 20 attention heads across 12 linguistic phenomena. Each head is ranked based on its performance for an individual phenomenon.

attention heads yield similar performance across different linguistic phenomena. Again, their performance is notably weaker in morphology but better for syntax and the semantics-syntax interface. While the performance of individual heads appears consistent across phenomena, certain heads (e.g., No.8 and No.16, as indicated by the ranking) consistently contribute the most. However, many heads display varied contributions across different phenomena, suggesting that each head might focus on distinct linguistic facets.

5. Discussion

The performance between models, and across individual layers of GPT2-XL offers a granular window into the linguistic representations formed by these language models.

Comparison between GloVe, ELMo, GPT The F1 scores landscape drawn from three models offers an intricate understanding of their competencies in Figure 2. The panels delineating morphology suggest that capturing word structures might be a hurdle for ELMo, as it lags behind GPT-2 XL. Similarly, in the semantics-syntax interface, ELMo's performance indicates potential challenges in grappling with the nuances between meaning and structure. In stark contrast, its performance in syntax showcases its adeptness in capturing abstract aspects of sentence structure. GloVe's unexpected strength in specific linguistic tasks, in spite of its inability to capture structural aspects of sentences, implies that strong baselines are necessary to isolate performance aspects of more complicated models that support most abstract linguistic representations. GPT-2 XL's relatively high performance across all linguistic phenomena, from morphology to syntax, emphasizes its comprehensive linguistic capabilities. It highlights not only the versatility of the model but also its relatively strong capacity to capture certain aspects of language structure.

GPT-2 XL's linguistic architecture Neural networks, particularly deep ones like GPT-2 XL, are known to capture features hierarchically. In the context of language modeling, simpler linguistic features, such as basic syntax and common word relationships, are often captured in the earlier layers. More abstract and complex features, like nuanced semantics and intricate grammatical relationships, tend to be represented in the middle to later layers (Yosinski et al., 2015). Thus, we suggest that the gradual accumulation of grammatical information through the first third of the layers might reflect this hierarchical capture of linguistic structures in Figure 3.

As we move deeper into the network, the model not only captures new information but also refines and contextualizes the features from the preceding layers. This could lead to some redundancy where the same grammatical features are represented across multiple layers (Raghu et al., 2017). Our results support this as we see relatively high performance across all later layers in the model, albeit the slight downward trend in Figure 3. As the model refines and merges features, some of the explicit grammatical information captured in the earlier layers might become implicit or get overshadowed by more complex linguistic patterns.

Morphology and semantics harder than syntax In essence, across models, performance thresholds, and the relationship between complexity and depth, all observations support specific hypotheses: the semantics-syntax interface and morphology present greater learning challenges for language models than pure syntax. This consensus aligns with previous studies suggesting that semantics is harder to learn than syntax (Tenney et al., 2019). Intriguingly, our results also align with neurolinguistic studies utilizing neural LMs to study the hierarchical mechanism of language processing. We speculate that the additional layers we observe for these pairs of sentences may be linked to the increased persistence of predictive repre-

sentations for semantic information as reported by Caucheteux et al. (2023), suggesting that semantic processing in the human brain is more long-term, high-level, than the short-term, shallow syntactic representation.

Attention's behavior Our findings on attention mechanisms in GPT-2 XL raise intriguing questions about their role and functionality. The absence of a clear incremental trend in attention, unlike hidden states, suggests that each attention layer in GPT-2 XL might be capturing unique linguistic information without necessarily building upon previous layers. The relatively low performance of attention heads in capturing morphology, especially given the supervised nature of our probing, might indicate that attention in GPT-2 XL isn't as adept at discerning morphological nuances.

The variability in performance among different attention heads may suggest an inherent specialization within the model. For instance, certain heads, like No.8 and No.16, consistently outperforming others could indicate that these heads have specialized in capturing more general or prevalent linguistic features. This could be a result of the training process, where frequent patterns in the data are more likely to be captured and optimized by specific heads.

6. Related Work

Probing LM Tenney et al. (2019) explored how much sentence structure is captured by contextually-embedded word representations. They used a series of diagnostic tasks to test whether these models can capture various grammatical and semantic features of sentences. What should be mentioned is while they did examine the outputs of LMs, their approach differed from a true per-layer probing in our study as they combined information from multiple layers so that it does not allow for a direct examination of specific linguistic information present within each layer.

Similarly, Hewitt and Manning (2019) delved into the intricacies of neural representations by introducing a "Structural Probe". This tool specifically quantifies the degree to which word representations in a sentence capture syntactic tree distances. Hewitt and Manning's approach highlights the potential of these embeddings to encapsulate the underlying syntactic structures of sentences, suggesting that neural network architectures inherently learn syntactic relationships as a consequence of their training. Manning et al. (2020) further offers intriguing insights into the emergent properties of neural networks. Their study underscored that even in the absence of explicit supervision or task-specific training signals, neural architectures can spontaneously develop internal representations that echo linguistic structures.

While these studies have made significant contributions to understanding the linguistic capabilities of intermediate embeddings, they often rely on broader linguistic tests or syntactic trees, which might not offer the granularity needed to dissect the specific linguistic phenomena being captured. The use of minimal pairs provides a finer-grained lens. It allows us to pinpoint with greater precision the exact linguistic distinctions these embeddings can discern, thereby offering a more detailed understanding of the nuances in linguistic representation across different layers.

Brain-inspired LM probing Concurrently, cognitive neuroscientists also provide insights from neuroscience into the internal neural representation of LMs. For instance, Caucheteux et al. (2021) delved into the GPT-2 embeddings and successfully isolated syntactic and semantic representations. Their findings further showed that these disentangled linguistic embeddings have cognitive neuroscience support, reinforcing the connection between neural LMs and language processing in the brain. While by using electrophysiology recording as a form of 'human measurement filter', Chen et al. (2023) suggest that intermediate layers of the deep speech model and language model share high-level contextual information. The present effort adds specificity to the kinds of information available to the network at particular layers, which in turn contributes to understanding how those layer-wise representations may, or may not, map to human neural states.

7. Conclusion and Future work

Adapting the decoding theory from cognitive neuroscience to language models, as we do in our study, offers a novel approach to understanding the internal mechanisms of these models. By treating the neural language model as a "brain", we can decode its internal representations, much like how cognitive neuroscientists decode brain activity. This methodological borrowing not only bridges the gap between cognitive neuroscience and natural language processing but also paves the way for new insights into the intricacies of linguistic representation within deep learning models.

Using this method, our analysis of GPT-2 XL, ELMo, and GloVe's capabilities across various grammaticality tasks sheds light on the inherent strengths and limitations of each model. The hierarchically layered architecture of deep neural models like GPT-2 XL captures linguistic features progressively, with simpler ones appearing in the early layers and more complex ones in subsequent layers. Consistently, our findings emphasize that the semantics-syntax interface and morphology are more challenging for LMs compared to syntax,

for both embeddings and attentions. In addition, results on individual attention heads show possible inherent specialization mechanisms.

Our current research focuses mainly on structure grammaticality, with limited attention to the conceptual capabilities of neural LMs. For future work, we aim to apply decoding probe methods to reveal LM's ability to master concepts using the COMPS minimal pair dataset (Misra et al., 2023), which is designed to probe such conceptual capabilities.

8. Bibliographical References

Tom B Brown, Benjamin Mann, Nick Ryder, Melvin Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pages 1336–1348. PMLR.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Peili Chen, Linyang He, Li Fu, Lu Fan, Edward F. Chang, and Yuanning Li. 2023. Do self-supervised speech and language models extract similar representations as human brain?

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

John-Dylan Haynes and Geraint Rees. 2006. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Junfeng Lu, Jinsong Wu, and Edward F Chang. 2022. Dissecting neural computations of the human auditory pathway using deep neural networks for speech. *bioRxiv*, pages 2022–03.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. 2006. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Openai blog*, 1(8).

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. In *International Conference on Machine Learning*, pages 821–829.

9. Language Resource References