Linguistically Grounded Analysis of Language Models using Shapley Head Values

Marcell Fekete
Department of Computer Science
Aalborg University
Copenhagen, Denmark
mrfe@cs.aau.dk

Johannes Bjerva
Department of Computer Science
Aalborg University
Copenhagen, Denmark
jbjerva@cs.aau.dk

Abstract

Understanding how linguistic knowledge is encoded in language models is crucial for improving their generalisation capabilities. In this paper, we investigate the processing of morphosyntactic phenomena, by leveraging a recently proposed method for probing language models via Shapley Head Values (SHVs). Using the English language BLiMP dataset, we test our approach on two widely used models, BERT and RoBERTa, and compare how linguistic constructions such as anaphor agreement and filler-gap dependencies are handled. Through quantitative pruning and qualitative clustering analysis, we demonstrate that attention heads responsible for processing related linguistic phenomena cluster together. Our results show that SHV-based attributions reveal distinct patterns across both models, providing insights into how language models organize and process linguistic information. These findings support the hypothesis that language models learn subnetworks corresponding to linguistic theory, with potential implications for cross-linguistic model analysis and interpretability in natural language processing (NLP).

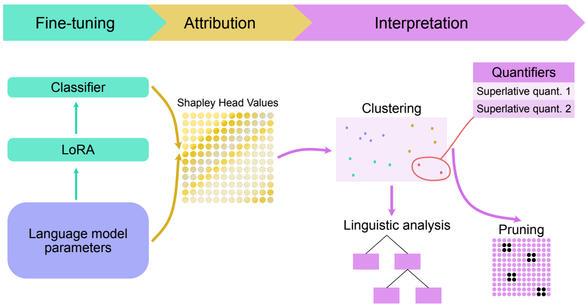

Figure 1: In the fine-tuning step, we train a classifier on a grammaticality judgement task. We carry out Shapley Head Value (SHV) attributions, and in the interpretation step, we carry out quantitative analysis using pruning as well as qualitative experiments using linguistic grounding.

1 Introduction

Language models gain knowledge of grammatical phenomena during pretraining. However, exactly how this knowledge is encoded is not well understood. While there is prior research on probing language models for morphosyntactic constructions (Finlayson et al., 2021; Mueller et al., 2022; Stanczak et al., 2022; Ács et al., 2023), it is not well established if this information is crucial to the model itself, or if it is merely learned as a by-product. We perform extensive analysis and offer evidence for a hypothesis that language models learn separate subnetworks which we can ground in linguistic theory, and are crucial to the model processing. To our knowledge this is the first paper that uses quantitative as well as qualitative analysis grounded in linguistic theory in assessing which language model components are responsible for taking care of the processing of specific linguistic phenomena.

We calculate Shapley Head Values (SHVs) following the methodology of Held and Yang (2023). While they use SHVs to find attention heads that have a high contribution to certain NLP tasks such as natural language inference to improve performance, we calculate SHVs using a grammaticality classification task across 13 different phenomena and 67 different constructions from the BLiMP benchmark (Warstadt et al., 2020, see Table 1). Hence, we offer a novel, linguistically grounded approach for isolating components of language models responsible for processing *specific* morphosyntactic phenomena. We cluster constructions based on SHVs, and assess the success of isolating heads responsible for processing aspects of morphosyntax using pruning based on relative importance of attention heads and linguistic analysis (see Figure 1).

In this paper, we perform an in-depth analysis using the BLiMP dataset to explore how widely used language models, such as BERT and RoBERTa,

Findings of the Association for Computational Linguistics: NAACL 2025, pages 850–865
April 29 - May 4, 2025

handle diverse morphosyntactic phenomena, including anaphor agreement, filler-gap dependencies, and island effects. Through a combination of quantitative pruning and qualitative linguistic analysis, we demonstrate that certain attention heads within these models are systematically responsible for processing distinct linguistic phenomena. Crucially, we show that related phenomena often cluster together within the models, suggesting that language models develop internal subnetworks corresponding to theoretical linguistic categories.

We make the following contributions:

1. We use Shapley Head Values (SHVs) to probe attention heads for their role in processing morphosyntactic phenomena, across two language models.

2. We apply SHV-based clustering, showing that language models develop subnetworks corresponding to related linguistic categories.

3. We provide qualitative linguistic analyses explaining the clustering of morphosyntactic phenomena based on SHV attributions.

4. Quantitative pruning validates our clusters by showing localized performance impacts when relevant attention heads are removed.

5. We release an implementation of our SHV-based probing and pruning methods.

2 Background

2.1 Probing

Previous work in attribution focuses on identifying the role that either training features or model components play with respect to various phenomena. *Training data attribution* is used to investigate, e.g., to what extent language-specific subnetworks rely on in-language examples from the training data when making predictions (Choenni et al., 2023). Foroutan et al. (2022) uses *iterative magnitude pruning* (Frankle and Carbin, 2019) to isolate sub-networks specific to languages and tasks like masked language modelling, named entity recognition, and natural language inference. *Structured pruning* is used for identifying attention heads important for dialogue summarisation (Liu and Chen, 2023) and cross-lingual natural language inference, e.g., with Shapley Head Values, (Held and Yang, 2023). *Extrinsic probing* efforts such as from Ács et al. (2023) target whether morphological information is encoded by language models. On the other hand, work in *intrinsic probing* aims to reveal how exactly linguistic information is structured within a model (Dalvi et al., 2019; Torroba Hennigen et al., 2020; Stanczak et al., 2022).

Another popular technique is *causal mediation analysis* (CMA), a causal method of identifying the importance of a model component to a target phenomenon (Pearl, 2001). Under CMA, two effects on the prediction of a model are compared: a direct effect of input intervention – such as a text edit – and an indirect effect. The relative degree of the indirect effect compared to the direct effect reveals the significance of the target component to the target phenomenon. CMA is used to investigate the role of individual neurons and attention heads in mediating gender bias (Vig et al., 2020), and to isolate neurons responsible for subject-verb agreement in English (Finlayson et al., 2021), and across a number of other languages (Mueller et al., 2022). A limitation of CMA, however, is that text edit operations are necessarily word-level to isolate other factors such as sensitivity to specific syntactic features. This makes it challenging to accommodate a number of morphosyntactic constructions such as island effects (see Table 1). Instead, we use SHVs which are particularly well-suited to probe for subnetworks responsible for processing of morphosyntactic phenomena.

Ours is certainly not the first paper to carry out the analysis of language model skills using linguistic grounding, but we believe our work is differentiated by the wider coverage in terms of morphosyntactic phenomena, as well as our efforts to localise linguistic knowledge. Previous work explores the relation between self-attention of input tokens and dependency links (Htut et al., 2019; Clark et al., 2019). Linzen and Baroni (2021) analyse the capabilities of LSTM and GRU models on long-distance dependencies, while Wilcox et al. (2024) use GPT-2 and GPT-3 to look at long-distance dependencies and island constraints.

2.2 Shapley Head Values

Shapley Values originate from game theory, devised to fairly distribute a given reward among a set of players based on their relative contribution to a certain outcome (Shapley, 1953; Mosca et al., 2022). They are used in model interpretability research thanks to their properties as attribution meth-

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	Many girls insulted **themselves**.	*Many girls insulted **herself**.
ARG. STRUCTURE	7	Rose wasn't **disturbing** Mark.	*Rose wasn't **boasting** Mark.
BINDING THEORY	7	Carlos said that Lori helped **him**.	*Carlos said that Lori helped **himself**.
CONTROL/RAISING	5	There was **bound** to be a fish escaping.	*There was **unable** to be a fish escaping.
DET.-NOUN AGR.	8	Rachelle had bought that **chair**.	*Rachelle had bought that **chairs**.
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**.
FILLER-GAP DEP.	7	Brett knew **what** many waiters find.	*Brett knew **that** many waiters find.
IRREG. FORMS	2	Aaron **broke** the unicycle.	*Aaron **broken** the unicycle.
ISLAND EFFECTS	8	Whose **hat** should Tonya wear?	*Whose should Tonya wear **hat**?
NPI LICENSING	7	The truck has **clearly** tipped over.	*The truck has **ever** tipped over.
QUANTIFIERS	4	No boy knew **fewer than** six guys.	*No boy knew **at most** six guys.
S-SELECTION	2	**Carrie** isn't listening to Jodi.	***That movie theater** isn't listening to Jodi.
SUBJ.-VERB AGR.	6	These casseroles **disgust** Kayla.	*These casseroles **disgusts** Kayla.

Table 1: Minimal pairs from the thirteen linguistic phenomena BLiMP paradigms are categorised into with the number of paradigms in each category (N). Differences are in **bold text** and, following convention, ungrammatical sentences are marked with an asterisk. Table adapted from Warstadt et al. (2020).

ods and satisfying theoretical properties of local accuracy, missingness and consistency (Shapley, 1953; Ghorbani and Zou, 2020; Held and Yang, 2023). Shapley Values have the advantage over alternative, gradient-based attribution methods in that they do not need evaluation functions to be differentiable, allowing them to be applied directly. They are also meaningfully signed with positive values reflecting positive contribution of the component and negative values reflecting the opposite (Held and Yang, 2023).. We use Shapley Head Values (SHVs) to measure the mean marginal contributions of attention heads in a language model in a linguistically grounded scenario, approximating these values following Held and Yang (2023). Concretely, we cluster morphosyntactic phenomena based on the similarity of their associated SHVs, then analyse the resulting cluster with regards to their correspondence to linguistic theory.

2.3 Pruning

According to the Lottery Ticket Hypothesis (Frankle and Carbin, 2019), neural models contain both harmful and beneficial connections between model components with respect to a target scenario. This means that pruning – the removing or turning off individual neurons or attention heads – can help us isolate subnetworks ('winning tickets') within language models (Pfeiffer et al., 2024). The common technique of pruning is to stop signals from passing through specific model components using a binary mask. Where the goal is improving model performance, the mask will impact components that contribute negatively to the target task, language, or other use cases. Pruning may also be utilised to discern how localised processing of a morphosyntactic phenomenon is in the model, and how generalisable this ability is to other phenomena. In our work, pruning is thus a quantitative metric to evaluate the cohesion of SHV clusters and the success of isolating components of a subnetwork encoding the same aspects of linguistic knowledge.

3 Methodology

3.1 Deriving SHVs

Following Held and Yang (2023), we define SHV φ_h, for a single attention head $\text{Att}_h \in A$[1], to represent the mean performance improvement on the characteristic function V – as derived from performance on the evaluation metric described in Section 3.2.2 – if Att_h contributes to the inference. To be able to remove or add attention heads at will, we augment our target models with a gate $G_h = \{0, 1\}$ for each attention head. When $G_h = 0$, the head Att_h is removed from the inference and does not contribute to the output of the transformer. The derivation of φ_h requires the contribution of the head Att_h to be measured across all Q permutations of the other gates, see Equation 1.

$$\varphi_h = \frac{1}{|Q|} \sum_{A \in Q} V(A \cup h) - V(A) \qquad (1)$$

[1] Representing the set of all attention heads.

Calculating Equation 1 for all of N attention heads requires 2^N evaluations, which is intractable with the number of heads language models contain. We can facilitate the computation through a number of steps (Ghorbani and Zou, 2020; Held and Yang, 2023). First, we can replace the full permutation set Q in Equation 1 by a randomly sampled subset of permutations via Monte Carlo simulations (Castro et al., 2009). Additionally, since SHV estimates are low-variance and computationally expensive to derive, we can speed up convergence by applying stopping criteria. One of these criteria is a truncation heuristic that determines stopping the sampling of the marginal contributions of a head once $< 50\%$ of attention heads remain in the permutation (Held and Yang, 2023).

The other stopping criterion is rooted in multi-armed bandit sampling. We want to stop sampling the marginal contributions of a head Att_h when we reach a decrease in the variance range in the approximated $\hat{\varphi}_h$ of Att_h. A low variance range indicates that we can be fairly confident in the degree of impact Att_h has for the characteristic function V. Our confidence interval is derived by Empirical Bernstein Bounds variance estimation (Maurer and Pontil, 2009). Given t samples with observed variance σ_t and a maximum variance range of R, the difference between the observed mean $\hat{\mu}$ and the true mean μ falls in range given by Equation 2 (Mnih et al., 2008) with a probability of $1 - \delta$.

$$|\hat{\mu} - \mu| \leq \sigma_t \sqrt{\frac{2\log(3/\delta)}{t}} + \frac{3R\log(3/\delta)}{t} \quad (2)$$

Sampling stops when the bound is less than $|\mu - 0|$ as this identifies the SHV as positive or negative with probability $1 - \delta$, which is an efficient way to determine which heads contribute positively to the target task. Following Held and Yang (2023), we set $R = 1$ and $\delta = 0.1$.

3.2 Experimental Setup

In the following, we introduce our experimental setup, including the dataset we use, the target task to derive SHVs, and our fine-tuning setup.

3.2.1 BLiMP Dataset

The Benchmark of Linguistic Minimal Pairs (BLiMP) is a challenge set for English that aims to measure the amount of linguistic knowledge language models have on a range of morphosyntactic phenomena (Warstadt et al., 2020). Constructions are organised into 67 minimal pair paradigms each containing 1,000 sentence pairs, and the individual paradigms are organised into 13 larger categories (see Table 1).[2] While similar datasets exist for Chinese with 38 (Song et al., 2022), for Japanese with 39 (Someya and Oseki, 2023), and Russian with 45 individual paradigms (Taktasheva et al., 2024), BLiMP is the largest in its category.

3.2.2 Grammaticality Judgement Task

We define the evaluation metric mentioned in Section 3.1 for the SHV derivation as accuracy on a custom grammaticality judgement task: a classifier has to output a binary label signifying which element of a sentence pair (s_1, s_2) is grammatical. The same metric is used to derive pruning performance. Both sentences are drawn from the same BLiMP paradigm p, member of the set of all paradigms $\mathcal{P}$, p shuffled prior to the selection of (s_1, s_2). Shuffling is done to facilitate the classifier focusing on the underlying morphosyntactic phenomenon represented – or violated – in the sentences, as opposed to merely the surface features of (s_1, s_2).

The performance of the classifier is measured in terms of accuracy, and it assigns label 0 or 1, depending on if s_1 or s_2 is grammatical:

$$\text{BinaryClass}(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 \text{ is grammatical,} \\ 1 & \text{if } s_2 \text{ is grammatical.} \end{cases}$$

The advantage of this novel task formulation is that – unlike other approaches such as simple text edits – it can flexibly incorporate morphosyntactic phenomena where the grammatical and ungrammatical sentences differ in more than a single word (see e.g., ELLIPSIS or ISLAND EFFECTS in Table 1).

3.2.3 Fine-Tuning Setup

We carry out our experiments and our analysis using the monolingual English models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). BERT has 110 million parameters and was trained on 16GB of data, while RoBERTa has 125 million parameters and was trained on 160GB. We follow Held and Yang (2023) in deriving SHVs (see Section 3.1). However, rather than fully fine-tuning our target model weights for the target task, we use

[2]For a full list of paradigms, see Appendix A.

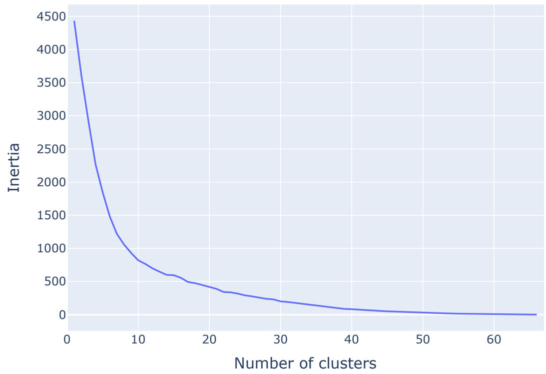

Figure 2: Clustering is done to try to optimise inertia and cluster count, resulting in an attempt at 10 clusters.

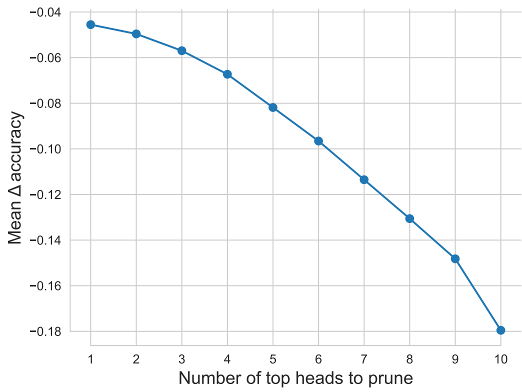

Figure 3: Mean Δ accuracy values drop in a near-linear fashion when pruning up to the top n heads across paradigms

low-rank adapters (LoRA; Hu et al. 2021) implemented with the `PEFT` library.[3] The main advantages of LoRA modules are their speed of training, and, more importantly, that they merely emphasise information that is already present in the original weights as opposed to reshaping the original model weights (Hu et al., 2021). This has the added advantage that it enables us to isolate and evaluate the pretrained linguistic knowledge of the models we consider.

We merge a subset of the sentence pairs from the individual BLiMP paradigms $p \in \mathcal{P}$ to create the training set for training LoRA and the classifier, and merge a smaller subset of pairs as a development set. For each p, 800 pairs go to the training set, and 100 pairs are assigned to the development set. After splitting the sentence pairs – but before merging into the training and development sets – we permute the set of ungrammatical sentences to avoid exact minimal pairs in order to better focus on the underlying grammatical construction (see Section 3.2.2). Finally, we shuffle the order of grammatical and ungrammatical sentences to create appropriate binary training examples, and merge the selections into the training and development sets. We retain the remaining 100 sentence pairs per paradigm to derive paradigm-specific attributions.

3.3 Interpretation

In this section, we describe the way we cluster BLiMP paradigms using SHVs, followed by how we interpret these clusters both with qualitative and quantitative means to assess how successfully we can identify subnetworks encoding specific linguistic knowledge across BERT and RoBERTa.

SHVs represent the mean marginal contributions of each attention head to the performance on the grammaticality judgement task defined in Section 3.2.2. For each BLiMP paradigm $p \in \mathcal{P}$, these SHV attributions are represented as vectors $\mathbf{v}_p \in \mathbb{R}^d$ where d is 144, i.e., the count of all attention heads in BERT and RoBERTa. We scale these vectors and cluster them using k-means clustering, grouping paradigms together based on how similarly the individual attention heads contribute to processing the target constructions. We decide an optimal number of clusters k by calculating inertia (see Figure 2). Based on empirical results, we pick $k = 10$, a small enough number to also facilitate in-depth analysis.

The goal of our qualitative, linguistic analysis is to understand the potential links between diverse paradigms that are clustered together, thus validating how successfully we identified the subnetworks responsible for these constructions. BLiMP paradigms are assigned into one of twelve categories reflecting morphosyntactic phenomena (see Table 1). On the highest level, our qualitative analysis is guided by the category membership of the individual paradigms – clusters containing paradigms from the same BLiMP category are likely to be cohesive in terms of morphosyntactic phenomena they represent. Clusters may be cohesive, however, even where category membership is not homogeneous. Categories may represent aspects of some of the same morphosyntactic phenomena – like how both BINDING THEORY and ANAPHOR AGR. are

[3]See `https://huggingface.co/docs/peft/index`.

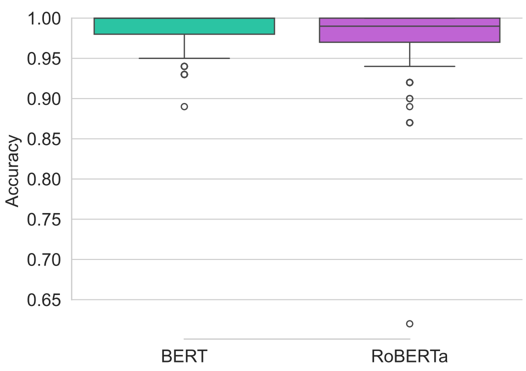

Figure 4: Distribution of baseline accuracy levels without pruning across the BERT and RoBERTa models.

concerned with the distribution of anaphors with respect to their antecedents. Category-level analysis can be complemented well by a sentence-level one.

Our quantitative analysis relies on pruning. We generate binary masks for each p paradigm, masking the top n attention heads in terms of their SHV, then apply this mask across all other paradigms $\mathcal{P}$. We do this to evaluate how cohesive the resulting clusters are, i.e., to what degree they represent that the language model has to apply some of the same set of morphosyntactic knowledge to process paradigms of the cluster. It stands to reason that if the cluster is well-defined, pruning using masks within cluster should have a larger impact than when utilising these masks for paradigms that are out-of-cluster. This is because out-of-cluster paradigms likely do not require the same set of morphosyntactic knowledge that is encoded in the relevant attention heads. As Figure 3 shows, pruning the top 10 heads already results in a large impact on accuracy across various paradigms. Since we target only 7% of all heads, we can observe how much of the processing of the various phenomena is localised as opposed to distributed more widely in the language model. This way we can verify how successfully we isolated the relevant components of a given subnetwork.

4 Results

We cluster the BLiMP paradigms into 10 individual clusters using the k-means algorithm based on the SHV vectors across BERT and RoBERTa. While

[4]There are two different **Binding** clusters in BERT, the paradigms in the smaller cluster are <u>underlined</u>.

	BLiMP Paradigm	Linguistics Term	M
NPI Cluster	NPI present (1)	NPI LICENSING	◊
	NPI present (2)		♦
	NPI scope ('only')		◆
	NPI scope (sentential negation)		◊
	NPI licensor present ('only')		◊
	" (matrix question)		◊
	" (sentential negation)		◊
	Irregular past participle verbs	IRREGULAR FORMS	♦
Island Effects	Adjunct island	ISLAND EFFECTS	◊
	Complex NP island		◊
	Coordinate structure constraint (complex left branch)		◊
	" (object extraction)		◆
	Left branch island (simple question)		◊
	Left branch island (echo question)		◊
	Wh-island		◆
Quantifiers	Superlative quantifiers 1	QUANTIFIERS	◊
	Superlative quantifiers 2		◊
Binding*	Anaphor gender agreement	ANAPHOR AGR.	◆
	Animate subject trans.	S-SELECTION	♦
	Principle A (case 1)	BINDING	♦
	" (domain 1)		◊
	" <u>(domain 2)</u>		◊
	" <u>(domain 3)</u>		◊
	" (c-command)		◊
Filler-Gap	Ellipsis N-bar (2)	ELLIPSIS	♦
	Existential 'there' (subject raising)	CONTROL/RAISING	♦
	Tough vs raising (2)		◆
	Principle A (case 1)	BINDING	◆
	" (case 2)		◊
	" (reconstruction)		◊
	Wh-questions (object gap)	FILLER-GAP DEP.	◊
	" (subject gap)		◊
	" (subject gap, long distance)		◊
	Wh vs 'that' (no gap)		◊
	" (no gap, long distance)		◊
Wh vs 'That'	Inchoative	ARG. STRUCTURE	◆
	Intransitive		◆
	Tough vs raising (1)	CONTROL/RAISING	◆
	Wh vs 'that' (with gap)	FILLER-GAP DEP.	◊
	" (with gap, long distance)		◊

Table 2: Clusters that emerge in both models with shared paradigms marked with ◊. A subset of paradigms only appear in in BERT (♦) or RoBERTa (◆). Clusters are named by the authors based on majority membership.[4]

	BLiMP Paradigm	Linguistics Term	M
Det.-Noun Cluster	Determiner-noun agr. (1)	DET.-NOUN AGR.	◆
	" (2)		◆
	" (irregular, 1)		◆
	" (irregular, 2)		◆
	" (with adjective, 1)		◆
	" (with adjective, 2)		◆
	" (with adjective, irregular 1)		◆
	" (with adjective, irregular 2)		◆
	Distractor agreement (relational noun)	SUBJECT-VERB AGR.	◆
	Regular plural subject-verb agr. (1)		◆
	Regular plural subject-verb agr. (2)		◆
	Transitive	ARG. STRUCTURE	◆

Table 3: Paradigms in the **Det.-Noun Cluster** in RoBERTa (◆).

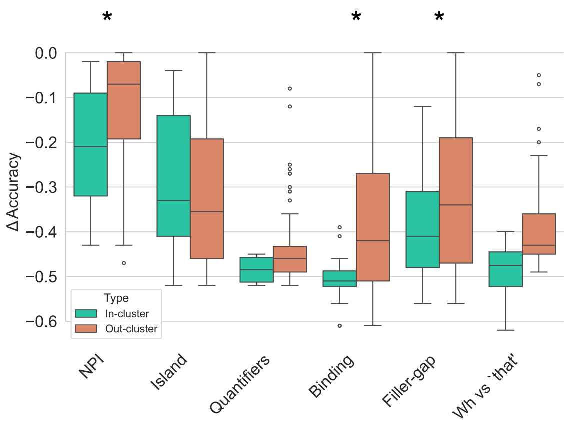

Figure 5: Impact in terms of Δ in accuracy in-cluster versus out-of-cluster across six BERT clusters when pruning the top 10 attention heads. Asterisks (*) show where the difference in distribution of the delta values is significant at $\alpha \leq 0.001$ after applying Bonferroni correction (Dunn, 1961) (see Appendix D).

clustering is stochastic, independent runs of the algorithm do not seem to differ significantly from one another.[5] As Table 2 demonstrates, six of the ten clusters show significant correspondence across the two language models. Some clusters, however, are less generalisable across the two models. The RoBERTa cluster in Table 3 contains paradigms representing a number of diverse linguistic phenomena, and the BERT cluster featuring DET.-NOUN AGREEMENT paradigms contains an even wider range of constructions.[6]

Figure 5 compares the relative impact of pruning in- and out-of-cluster across the BERT clusters in Table 2, i.e., the ones that have related RoBERTa clusters. This impact is measured in the change (Δ) from the baseline values in terms of accuracy on the grammaticality judgement task when pruning the top 10 attention heads. When in-cluster pruning in general results in a more dramatic negative Δ than pruning out-of-cluster, we can surmise that the top 10 attention heads contribute significantly to the processing of the target paradigms, thus we successfully identified the most important parts of a subnetwork. In other cases, e.g., the **Island** or **Quantifiers** cluster, either the target phenomenon is not localised enough for the top 10 heads to cause a significant impact, or the clusters are not homogeneous or exclusive enough with respect to the phenomena they contain.

[5]See Appendix B for analysis.

[6]See all clusters in Appendix C.

5 Discussion

In the following, we discuss our most important findings regarding the clusters created via SHV attributions. The membership of these clusters indicates which morphosyntactic phenomena are processed using the same subnetworks by BERT and RoBERTa, thus showing how the models generalise over these constructions.

Consistency across models Cluster membership across the BERT and RoBERTa models largely corresponds between the majority – 6 out of 10 – of the clusters (see Table 2).

Successful grouping of linguistic categories In many cases, BLiMP paradigms from the same linguistic categories appear in their own clusters. This is especially true for the clusters that are consistent across the language models. For instance, NPI LICENSING or BINDING paradigms mostly appear in their **NPI** or **Binding** cluster, respectively (see Table 2). DET.-NOUN AGR. paradigms are in their own – though less homogeneous – cluster in the RoBERTa model (Table 3) and the BERT model as well (see Table 7 in Appendix C).

Common ground across linguistic categories While clusters do tend to collect BLiMP paradigms from the same categories, exceptions, i.e., the presence of paradigms from other linguistic categories, can often be explained by linguistic analysis. Take the **Filler-gap** cluster (Table 2). Paradigms in FILLER-GAP DEP. typically represent the fronting of linguistic material that can be analysed to have originated in a different clause (1). Similarly, relevant CONTROL/RAISING (2) and BINDING (3) paradigms also deal with the licensing of raised material. Finally, ELLIPSIS can be analysed as a link between an antecedent and consequent clause which allows the omission of linguistic material (e in 4).[7]

(1) Wayne has revealed who/*that most hospitals admired t.

(2) There was bound/*unable to be a fish escaping.

(3) It's himself that this cashier attacked t /*It's himself that attacked this cashier.

[7]The symbol t for trace is used as a convention to indicate where the raised material originates from, while e represents the omitted material.

(4) This print scares a lot of busy senators and Benjamin scares a few *e* /*This print scares a lot of senators and Benjamin scares a few busy.

The **Binding** cluster contains BLiMP paradigms related to phenomena concerning anaphors. BINDING paradigms represent the various licensing restrictions of anaphors which is often represented by the presence or lack of gender agreement between the anaphor and the antecedent noun (5), similarly to the relevant ANAPHOR AGR. paradigm (6).

(5) Gina explains Alan fires himself/*Alan explains Gina fires himself.

(6) A girl couldn't reveal herself/*himself.

Finally, the DET.-NOUN AGR. and SUBJECT-VERB AGR. paradigms in the RoBERTa cluster in Table 3 are connected by the fact that both phenomena concern number agreement (7a and 7b).

(7) a. Raymond is selling this sketch/*sketches.

b. The students/*student perform.

Linguistic knowledge is often strongly localised The difference between in-cluster and out-of-cluster pruning impact on Δ accuracy indicates we identified the majority of attention heads representing the full construction-specific subnetwork (see Figure 5). In a subset of cases this is not the case: either the subnetwork is larger, or the relevant knowledge is more widely distributed across model components. Three out of the six clusters show a significant difference between in-cluster and out-of-cluster Δ accuracy. This is very unlikely to occur randomly: in our experiments with 125 random clusters of varying sizes, we found it to happen only in two cases.

Different degrees of sensitivity between models SHVs prove useful to cluster related BLiMP paradigms into clusters, but it is clear that the RoBERTa model is somewhat more discerning with regards to morphosyntactic phenomena. In the case of BERT, 25 of all 67 paradigms are assigned to a single cluster representing 9 different linguistic categories (see **BERT 3** in Table 7, Appendix C). RoBERTa has no such cluster that collects this number of unrelated linguistics constructions. The discrepancy between the two language models may lie in the comparably higher degree of linguistic knowledge obtained by RoBERTa thanks to the fact that it is pretrained on vastly more data than BERT for more training steps, and that it has more model parameters (see Section 3.2.3). This is also shown how its performance surpasses BERT on many metrics (Liu et al., 2019).

These points show that language models encode *at least* a subset of all morphosyntactic knowledge in subnetworks that our methodology can identify. This enables us to evaluate the generalisation abilities of language models between related linguistic constructions. We can additionally find that the success of generalisation is impacted by the depth of pretraining, i.e., the size of pretraining data and the number of training steps.

6 Conclusion

We apply intrinsic probing on two commonly used language models, BERT and RoBERTa, in order to investigate how linguistic knowledge is represented and organised internally. We show that attributions based on SHVs can be used to identify attention heads of subnetworks that generalise across related morphosyntactic phenomena, and allow us to carry out a linguistically grounded analysis. In this, we showcase that SHVs are well-suited to provide linguistically interpretable insights into the inner workings of language models, beyond the task-specific investigations carried out by Held and Yang (2023). Additionally, our pruning analysis demonstrates that in many cases, the identified attention heads are crucial components of the identified subnetworks; we find that switching these attention heads off severely impacts the grammaticality judgement of the language models. In future work, our methods can prove valuable in describing subnetworks specific to linguistic knowledge in language models. This might be particularly valuable in cross-lingual settings.

Limitations

A major limitation for our work is the difficulty of scaling it to new languages and to new language models. First, the coverage of datasets of linguistic minimal pairs for languages other than English lags behind that of BLiMP. Next, linguistically grounded clustering analysis requires an in-depth understanding of the specific morphosyntactic phenomena in a language, as well as considerable manual effort. This effort is a bottleneck on how readily the analysis can be carried out for other language models and, of course, other languages. Additionally, sentences in the BLiMP dataset are generated from expert templates using a predefined vocabulary, which means they tend to be semantically empty as well as heavily formulaic. However, since this is true for both grammatical and ungrammatical sentences, this fact is not likely to cause problems for our attributions. Furthermore, the derivation of SHVs requires the use of sampling techniques and truncation heuristics to make computation tractable. These techniques introduce uncertainty, which means there is a slight chance they might influence the accuracy of the attribution process. Moreover, the binary gates used to zero attention head activations may harm model performance since these zero-values are out-of-distribution for the language model. This means the actual importance of attention heads somewhat difficult to discombobulate from the impact of the zero ablations. Finally, we are carrying out attributions on the level of the attention head, rather than on the level of the neuron. This improves tractability and facilitates the qualitative analysis, but it may result in a loss of granularity. In future work, it might be worth exploring what we can gain from neuron-level attributions.

Ethics Statement

This paper deals with a linguistically grounded analysis of the inner workings of language models. Our work adheres to the ACL ethics guidelines.

Acknowledgements

MF and JB are funded by the Carlsberg Foundation, under the *Semper Ardens: Accelerate* programme (project nr. CF21-0454).

References

Judit Ács, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, pages 1–40.

A Asudeh and M Dalrymple. 2006. Binding Theory. In *Encyploedia of Language & Linguistics*, second edition. Elsevier Ltd.

A Carnie. 2006. Island Constraints. In *Encyploedia of Language & Linguistics*, second edition. Elsevier Ltd.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Rochelle Choenni, Ekaterina Shutova, and Dan Garrette. 2023. Examining Modularity in Multilingual LMs via Language-Specialized Subnetworks. ArXiv:2311.08273 [cs].

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, Florence, Italy. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

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

S.W. Dubinsky and W.D. Davies. 2006. Control and Raising. In *Encyclopedia of Language & Linguistics*, second edition, pages 131–139. Elsevier Ltd.

Olive Jean Dunn. 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64.

Y N Falk. 2006. Long-Distance Dependencies. In *Encyclopedia of Language & Linguistics*, second edition. Elsevier Ltd.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. ArXiv:1803.03635 [cs].

Amirata Ghorbani and James Y Zou. 2020. Neuron Shapley: Discovering the Responsible Neurons. In *Advances in Neural Information Processing Systems*, volume 33, pages 5922–5932. Curran Associates, Inc.

William Held and Diyi Yang. 2023. Shapley head pruning: Identifying and removing interference in multilingual transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2416–2427, Dubrovnik, Croatia. Association for Computational Linguistics.

J Hoeksema. 2006. Polarity Items. In *Encyclopedia of Language & Linguistics*, second edition. Elsevier Ltd.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? ArXiv:1911.12246.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. ArXiv:2106.09685 [cs].

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(1):195–212.

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:1907.11692 [cs].

Zhengyuan Liu and Nancy F. Chen. 2023. Picking the Underused Heads: A Network Pruning Perspective of Attention Head Selection for Fusing Dialogue Coreference Information. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece. IEEE.

Andreas Maurer and Massimiliano Pontil. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. ArXiv:0907.3740 [stat].

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. 2008. Empirical Bernstein stopping. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 672–679, Helsinki, Finland. ACM Press.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Event-place: Seattle, Washington.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2024. Modular Deep Learning. ArXiv:2302.11529 [cs].

L. S. Shapley. 1953. 17. A Value for n-Person Games. In Harold William Kuhn and Albert William Tucker, editors, *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, Princeton.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs. ArXiv:2406.19232 [cs].

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In

Advances in Neural Information Processing Systems, volume 33, pages 12388–12401. Curran Associates, Inc.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

S Winkler. 2006. Ellipsis. In *Encyclopedia of Language & Linguistics*, second edition. Elsevier Ltd.

A All BLiMP Paradigms

See Table 4 for a list of all minimal pair paradigms and examples in the BLiMP dataset (Warstadt et al., 2020), organised into their relevant linguistic categories.

B Cluster Purity

Clusters	Purity: μ (σ)
BERT REFERENCE	
BERT clusters	0.497 (0.045)
RoBERTa clusters	0.530 (0.046)
Random clusters	0.277 (0.027)
ROBERTA REFERENCE	
BERT clusters	0.532 (0.044)
RoBERTa clusters	0.765 (0.063)
Random clusters	0.285 (0.028)

Table 5: Mean (μ) purity scores and standard deviations (σ) across k-means runs measured against our BERT and RoBERTa reference clusters. Each purity score represents comparison with 100 individual runs, within-model comparisons are underlined.

Since k-means clustering is a stochastic process, SHV clusters may differ between runs of the algorithm. We focus on showing how qualitative and quantitative analysis can reveal shared patterns between clusters in general, thus ensuring clustering consistency is not our goal. Nevertheless, we evaluate do carry out intrinsic evaluation of the method using *purity* as a metric. Given N data points, a set of clusters $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$, and a set of classes $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$, purity is calculated through assigning each cluster to the class most frequent in the cluster, and then measuring the number of correctly assigned data points divided by N, see Equation 3.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \quad (3)$$

Purity falls between 0 (no match between clusters) and 1 (perfect match), and it is typically calculated against a gold standard cluster set. As we do not have gold clusters, we measure the purity using reference clusters for both BERT and RoBERTa (see the clusters in Section 4). Since cluster labels may also change between runs, we cannot simply use cluster IDs as gold labels even using the reference clusters. Instead, we aim to align clusters

BLiMP Paradigm	Grammatical/Ungrammatical Examples
	Anaphor agr.
Anaphor gender agreement	A girl couldn't reveal herself/*himself.
Anaphor number agreement	Thomas complained about himself/*themselves.
	Arg. structure
Causative	Aaron breaks/*appeared the glass.
Drop argument	Travis is touring/*Travis is revealing.
Inchoative	Patricia had changed./*Patricia had forgotten.
Intransitive	The screen does brighten/*resemble.
Passive 1	Tracy isn't fired/*muttered by Jodi's daughter.
Passive 2	Steve isn't disliked/*lied.
Transitive	Diane watched/*screamed Alan.
	Binding
Principle A (c-command)	A girl that wouldn't watch Omar questions herself/*himself.
" (case 1)	The teenagers explain that they/*themselves aren't breaking all glasses.
" (case 2)	Eric imagines himself taking every rug/*Eric imagines himself took every rug.
" (domain 1)	Carla had explained that Samuel has discussed her/*herself.
" (domain 2)	James says Kayla helped herself/*himself.
" (domain 3)	Gina explains Alan fires himself/*Alan explains Gina fires himself.
" (reconstruction)	It's himself that this cashier attacked/*It's himself that attacked this cashier.
	Control/raising
Existential 'there' (object raising)	Frank judged /*compelled there to be a photograph of Michael looking like Sherry.
" (subject raising)	There was bound to be a fish escaping/*There was unable to be a fish escaping.
Expletive 'it' object raising	This cashier had ascertained/*can't press it to be not so interesting that Anna painted.
'Tough' vs raising (1)	Julia wasn't fun/*unlikely to talk to.
" (2)	Bruce was sure/*annoying to remember Gerald.
	Det.-noun agr.
Determiner-noun agreement (1)	Raymond is selling this sketch/*sketches.
" (2)	Tracy passed these/*this art galleries.
" (irregular 1)	The driver reveals these/*this mice.
" (irregular 2)	Natalie describes this/*these child.
" (with adjective 1)	Many men have these messy cups/*cup.
" (with adjective 2)	Donna might hire this/*these serious actress.
" (with adjective, irregular 1)	Heidi returns to that big woman/*women.
" (with adjective, irregular 2)	Denise did confuse that/*those important women.
	Ellipsis
Ellipsis N-bar (1)	This print scares a lot of busy senators and Benjamin scares a few/*This print scares a lot of senators and Benjamin scares a few busy.
" (2)	Vincent wore one shirt and Matt wore some big shirt/*Vincent wore one hidden shirt and Matt wore some big.
	Filler-gap dep.
Wh-questions (object gap)	Joel discovered the vase that Patricia might take/*Joel discovered what Patricia might take the vase.
" (subject gap)	Leslie remembered some guest that has bothered women./*Leslie remembered who some guest has bothered women.
" (subject gap, long distance)	Regina sees that candle that Steve lifts that might impress every doctor./*Regina sees who that candle that Steve lifts might impress every doctor.
Wh vs 'that' (no gap)	Mark figured out that/*who most governments appreciate Steven.
" (no gap, long distance)	Eva discovered that/*who all pedestrians that have performed upset Candice.
" (with gap)	Wayne has revealed who/*that most hospitals admired.
" (with gap, long distance)	Kenneth investigated who/*that the cashiers that perform cared for.
	Irregular forms
Irregular past participle (adjectives)	The broken/*broke mirrors were blue.
" (verbs)	The Borgias wore/*worn a lot of scarves.
	Island effects
Adjunct island	Who should Derek hug after shocking Richard?/*Who should Derek hug Richard after shocking?
Complex NP island	What can't a guest who would like some actor argue about?/*What can't some actor argue about a guest who would like?
Coordinate structure constraint (complex left branch)	Which teenagers had Tamara hired and Grace fired?/*Which hard Tamara hired teenagers and Grace fired?
" (object extraction)	What had Russel and Douglas attacked?/*What had Russel attacked and Douglas?
Left branch island (echo question)	Irene had messed up whose rug?/*Whose had Irene messed up rug?
" (simple question)	Whose museums had Dana alarmed?/*Whose had Dana alarmed museums?
Sentential subject island	Who should pedestrians' curing Deanna scare/*Who should pedestrians' curing scare Deanna.
Wh-island	Who isn't Craig realising he/*who kisses?
	NPI licensing
NPI licensor present (matrix question)	Had Bruce ever played? / *Bruce had ever played.
" ('only')	Only/*Even Bill would ever complain.
" (sentential negation)	Teresa had not/*probably every sold a movie theater.
NPI present (1)	Even Suzanne has really/*ever joked around.
" (2)	Tamara really/*ever exited these mountains.
NPI scope ('only')	Only many people who George likes ever clashed./*Many people who only George likes ever clashed.
" (sentential negation)	Every coat that did scare Nina has not ever wrinkled/*Every coat that did not scare Nina has ever wrinkled.
	Quantifiers
Existential 'there' quantifiers (1)	There was a/*each documentary about music irritating Allison.
" (2)	All dancers are there talking to Pamela./*There are all dancers talking to Pamela.
Superlative quantifiers (1)	No girl attacked fewer/*at most than two waiters.
" (2)	The/*No hospital had fired at most four people.
	S-selection
Animate subject (passive)	Lisa was kissed by the boys/*blouses.
" (transitive)	Phillip/*This pasta can talk to those waitresses.
	Subject-verb agr.
Distractor agreement (relational noun)	A story about the Balkans doesn't/*don't irritate a person.
" (relative clause)	Boys that aren't disturbing Natalie suffer/*suffers.
Irregular plural subject verb agreement (1)	This goose isn't/*weren't bothering Edward.
" (2)	The people/*person conspire.
Regular plural subject verb agreement (1)	The cups alarm/*alarms Angela.
" (2)	The students/*student perform.

Table 4: A list of all BLiMP minimal pair paradigms and examples, organised according to their respective linguistic categories.

using the Hungarian (or Kuhn-Munkres) algorithm that matches clusters by maximising shared datapoints between them.

Table 5 shows purity scores between BERT and RoBERTa clusters and three sets of 100 different cluster sets: these include other runs of k-means clustering on the BERT and RoBERTa SHVs, and randomly assigned clusters. It is clear that the cohesion of clusters as measured in terms of purity across and within models far exceeds the cohesion between the reference clusters and the random clusters.

C All Clusters

See Tables 6 and 7 that list all BERT and RoBERTa clusters, representing both matching and (underlined in the tables) and not matching ones.

D T-Test on Pruning In- and Out-of-Cluster

Cluster	T-stat.	P-value
NPI	−4.809	1.12e−05
Island	0.995	0.329
Quantifiers	−2.053	0.113
Binding	−8.719	7.97e−11
Filler-gap	−4.885	2.94e−06
Wh vs 'that'	−1.855	0.157

Table 8: Results of the T-test between in-cluster and out-of-cluster pruning with the BERT model.

Table 8 shows T-test statistics and P-values – adjusted using the Bonferroni correction (Dunn, 1961) – between accuracy changes in the in-cluster and out-of-cluster pruning of relevant clusters with the BERT model. The **NPI**, **Binding**, and **Filler-gap** clusters pass the T-test, i.e., the differences between the distribution of in- and out-of-cluster values are statistically significant. This indicates a more drastic consequence of pruning within BLiMP paradigms in these clusters than when using prune masks from paradigms outside of the clusters.

This is not the case in the other three clusters. However, both the **Quantifiers** and **Wh vs 'that'** clusters contain only 2-2 paradigms, meaning that in-cluster pruning involves only four datapoints. This makes the success of any statistical analysis questionable. Finally, the **Island** cluster contains only 5 out of 8 ISLAND EFFECTS paradigms. Additionally, it is also possible that the attention heads responsible for processing these paradigms are not so well localisable as with other paradigms. This may reduce the impact of pruning only the top 10 attention heads.

E Glossary of Linguistic Terms

In this section, we include a glossary of some linguistic terms relevant to BLiMP that merit definition.

Anaphors and binding theory Binding theory conditions the distribution of nominals, particularly pronouns and anaphors, i.e., reflexive pronouns (Asudeh and Dalrymple, 2006). The main constraint occurs with respect to a potential antecedent nominal that co-refers with the target nominal. In the following discussion, this co-reference is signalled by indexing with i and j. Anaphors can only occur if there is a valid antecedent in the sentence, see the example in (5) restated in (8):

(8) Gina$_i$ explains Alan$_j$ fires himself$_j$.

Pronouns can have a valid antecedent but not in the same clause as them:

(9) Gina$_i$ explains Alan$_i$ fires her$_i$.

Control and raising Both constructions involve a noun phrase (NP) in a main clause determining a covert reference of a subject of a subordinate clause. The key difference is that under raising this main clause NP (in M) receives its semantic role from the verb of the subordinate clause '**expected**' (in S), see in (10). Under control, on the other hand, the NP receives its semantic role from the verb of the main clause '**persuaded**' (11).

(10) [$_M$ The teacher **expected** <u>the students</u>]$_M$ [$_S$ to help the visitors]$_S$.

(11) [$_M$ The teacher **persuaded** <u>the students</u>]$_M$ [$_S$ to help the visitors]$_S$.

In practice, however, both constructions involve a subordinate that is seemingly lacking a subject (Dubinsky and Davies, 2006).

Ellipsis Ellipsis refers to the "omission of linguistic material, structure, and sound" (Winkler, 2006). Certainly ellipsis is not an unbounded phenomenon, and typically the omitted material stands in a co-reference relation with some overt element in the sentence.

	BLiMP Paradigm	Linguistics Term	M
NPI Cluster	NPI present (1)	NPI LICENSING	◊
	NPI present (2)		♦
	NPI scope (‘only’)		◆
	NPI scope (sentential negation)		◊
	NPI licensor present (‘only’)		◊
	" (matrix question)		◊
	" (sentential negation)		◊
	Irregular past participle verbs	IRREGULAR FORMS	♦
Island Effects	Adjunct island	ISLAND EFFECTS	◊
	Complex NP island		◊
	Coordinate structure constraint (complex left branch)		◊
	" (object extraction)		◆
	Left branch island (simple question)		◊
	Left branch island (echo question)		◊
	Wh-island		◆
Quan-tifiers	Superlative quantifiers 1	QUANTIFIERS	◊
	Superlative quantifiers 2		◊
Binding*	Anaphor gender agreement	ANAPHOR AGR.	◆
	Animate subject trans.	S-SELECTION	♦
	Principle A (case 1)	BINDING	♦
	" (domain 1)		◊
	" (domain 2)		◊
	" (domain 3)		◊
	" (c-command)		◊
Filler-Gap	Ellipsis N-bar (2)	ELLIPSIS	♦
	Existential ‘there’ (subject raising)	CONTROL/RAISING	♦
	Tough vs raising (2)		◆
	Principle A (case 1)	BINDING	◆
	" (case 2)		◊
	" (reconstruction)		◊
	Wh-questions (object gap)	FILLER-GAP DEP.	◊
	" (subject gap)		◊
	" (subject gap, long distance)		◊
	Wh vs ‘that’ (no gap)		◊
	" (no gap, long distance)		◊
Wh vs ‘That’	Inchoative	ARG. STRUCTURE	◆
	Intransitive		◆
	Tough vs raising (1)	CONTROL/RAISING	◆
	Wh vs ‘that’ (with gap)	FILLER-GAP DEP.	◊
	" (with gap, long distance)		◊
BERT 1	Anaphor gender agreement	ANAPHOR AGR.	♦
	Distractor agreement (relational noun)	SUBJECT-VERB AGR.	♦
	" (relative clause)		♦
	Irregular plural subject verb agreement (1)		♦
	Existential ‘there’ (object raising)	CONTROL/RAISING	♦
	Expletive ‘it’ (object raising)		♦
	NPI scope (‘only’)	NPI LICENSING	♦
	Sentential subject island	ISLAND EFFECTS	♦
BERT 2	Irregular past participle adjectives	IRREGULAR FORMS	♦
	Passive (1)	ARG. STRUCTURE	♦
	Passive (2)		♦
RoBERTa 1	Animate subject (passive)	S-SELECTION	◆
	" (transitive)		◆
	Passive (1)	ARG. STRUCTURE	◆

Table 6: BLiMP paradigms in the relevant BERT (♦) and RoBERTa (◆) clusters. Paradigms that appear in the clusters of both models are marked with ◊. Clusters without obvious organising principles are marked with model names and numbers instead of labels.

	BLiMP Paradigm	Linguistics Term	M
BERT 3	Anaphor number agreement	ANAPHOR AGR.	♦
	Animate subject (passive)	S-SELECTION	♦
	Causative	ARG. STRUCTURE	♦
	Drop argument		♦
	Inchoative		♦
	Intransitive		♦
	Transitive		♦
	Coordinate structure constraint (object extraction)	ISLAND EFFECTS	♦
	Wh-island		♦
	Determiner-noun agr. (1)	DET.-NOUN AGR.	♦
	" (2)		♦
	" (irregular, 1)		♦
	" (irregular, 2)		♦
	" (with adjective, 1)		♦
	" (with adjective, 2)		♦
	" (with adjective, irregular 1)		♦
	" (with adjective, irregular 2)		♦
	Ellipsis N-bar (1)	ELLIPSIS	♦
	Existential 'there' quantifiers (1)	QUANTIFIERS	♦
	" (2)		♦
	Irregular plural subject-verb agreement (2)	SUBJECT-VERB AGR.	♦
	Regular plural subject-verb agreement (1)		♦
	" (2)		♦
	'Tough' vs raising (1)	CONTROL/RAISING	♦
	" (2)		♦
Det.-Noun Cluster (RoBERTa)	Determiner-noun agr. (1)	DET.-NOUN AGR.	♦
	" (2)		♦
	" (irregular, 1)		♦
	" (irregular, 2)		♦
	" (with adjective, 1)		♦
	" (with adjective, 2)		♦
	" (with adjective, irregular 1)		♦
	" (with adjective, irregular 2)		♦
	Distractor agreement (relational noun)	SUBJECT-VERB AGR.	♦
	Regular plural subject-verb agr. (1)		♦
	Regular plural subject-verb agr. (2)		♦
	Transitive	ARG. STRUCTURE	♦
RoBERTa 2	Anaphor number agreement	ANAPHOR AGR.	♦
	Causative	ARG. STRUCTURE	♦
	Drop argument		♦
	Passive (2)		♦
	Ellipsis N-bar (1)	ELLIPSIS	♦
	" (2)		♦
	Irregular past participle (adjectives)	IRREGULAR FORMS	♦
	" (verbs)		♦
	Irregular plural subject-verb agreement (1)	SUBJECT-VERB AGR.	♦
	" (2)		♦
	NPI present (2)	NPI LICENSING	♦
	Sentential subject island	ISLAND EFFECTS	♦
RoBERTa 3	Distractor agreement (relative clause)	SUBJECT-VERB AGR.	♦
	Existential 'there' object raising	CONTROL/RAISING	♦
	Existential 'there' quantifiers (1)	QUANTIFIERS	♦
	" (2)		♦
	Existential 'there' subject raising	CONTROL/RAISING	♦
	Expletive 'it' object raising		♦

Table 7: BLiMP paradigms in the relevant BERT (♦) and RoBERTa (♦) clusters (continued). Paradigms that appear in the clusters of both models are marked with ♢. Clusters without obvious organising principles are marked with model names and numbers instead of labels.

Filler-gap dependencies Also called long-distance dependencies, 'filler' refers to a fronted element and the 'gap' is the position with which it is semantically or syntactically related (Falk, 2006). Examples for such dependencies involve *wh* questions, exclamatives, topicalisation, cleft, and more, see the examples in (12).

(12) a. Which painting does the artist believe the curator said the gallery owner hung t on the wall?

b. What a painting the artist believes the curator said the gallery owner hung t on the wall!

c. This painting, the artist believes the curator said the gallery owner hung t on the wall.

d. It is the painting that the artist believes the curator said the gallery owner hung t on the wall.

Island effects Under certain (generative) linguistic analyses, various elements of a sentence materialise in a different location than where they were generated in (Carnie, 2006). For instance, *wh*-movement, i.e., the displacement of a question word is a common example for this, see (13).

(13) a. Mary does love pineapples.

b. What does Mary love t?

The trace t indicates where the question word *What* originated from under the movement analysis. Islands, on the other hand, are a collection of constraints on movement. These typically contain restrictions on moving across various clauses. There are many such constraints, but one example is the so-called Coordinate Structure Constraint that disallows extracting members of a conjunction, see the examples in (14).

(14) a. I have eaten the salad and the pizza.

b. *What have you eaten the salad and t?

c. *What have you eaten t and the pizza?

NPI Negative Polarity Items (NPI) – like *any*, *ever*, or *yet* – are words that appear only in a limited number of contexts. These contexts are among all, the scope of negation, as complements of negative predicates, in comparative clauses, in questions, and in the scope of negative quantifiers and adverbs such as *few/little*, *rarely*, or *only* (Hoeksema, 2006).