Disjoint Processing Mechanisms of Hierarchical and Linear Grammars in Large Language Models

Anonymous ACL submission

Abstract

All natural languages are structured hierarchically. In humans, this structural restriction is neurologically coded: when two grammars are presented with identical vocabularies, brain areas responsible for language processing are only sensitive to hierarchical grammars. Using large language models (LLMs), we investigate whether such functionally distinct hierarchical processing regions can arise solely from exposure to large-scale language distributions. We generate inputs using English, Italian, Japanese, or nonce words, varying the underlying grammars to conform to either hierarchical or linear/positional rules. Using these grammars, we first observe that language models show distinct behaviors on hierarchical versus linearly structured inputs. Then, we find that the components responsible for processing hierarchical grammars are distinct from those that process linear grammars; we causally verify this in ablation experiments. Finally, we observe that hierarchy-selective components are also active on nonce grammars; this suggests that hierarchy sensitivity is not tied to meaning, nor in-distribution inputs.

1 Introduction

In 1861, Broca found evidence that language processing functions are *localized* in specific brain regions. Since then, our mapping of the brain has advanced tremendously; we now know that **functional specialization** can arise not only from biologically coded mechanisms, but also from experience (Baker et al., 2007). More recently, there has been significant interest in understanding the mechanisms of language processing in large language models (Olsson et al., 2022; Hanna et al., 2023; Yu et al., 2023; Todd et al., 2024), whose inductive biases are more general than those of humans.

Sensitivity and **functional selectivity** toward the hierarchical structure of language is a hallmark of human language processing (Chomsky, 1957, 1965). Hierarchical grammars follow the structure of natural language, where elements of a sentence are arranged according to syntactic rules and dependencies. These grammars reflect how languages are naturally processed by humans: they often incorporate recursion and create dependencies between words based on grammatical roles rather than their position in the sentence. In contrast, linear grammars arrange elements based on fixed positional rules or relative word ordering. Importantly, brain regions that are selective for hierarchical grammars are disjoint from those that process linear structures, as well as from those involved in hierarchical but non-linguistic domains such as music or programming languages, or sentences constructed from nonce words (Malik-Moraleda et al., 2023; Fedorenko et al., 2016; Ivanova et al., 2020; Liu et al., 2020; Varley and Siegal, 2000; Varley et al., 2005; Apperly et al., 2006; Fedorenko and Varley, 2016; Monti et al., 2009; Fedorenko et al., 2011; Amalric and Dehaene, 2019; Ivanova et al., 2021; Chen et al., 2023).

Despite large quantities of evidence from humans for language selectivity, it is not clear whether language models would *acquire* similar selectivity from exposure to natural language data in the absence of human-like learning biases. Kallini et al. (2024) recently find that autoregressive Transformer-based models (Vaswani et al., 2017) can more easily learn grammars that accord with the structures found in human language. While that study provides evidence from language acquisition, we are primarily interested in language processing in models trained on large text corpora.

Do large language models (LLMs) demonstrate distinct mechanisms for processing hierarchically structured vs. non-hierarchically structured sentences that are otherwise superficially identical? We derive inspiration from Musso et al.'s (2003) experiment testing hierarchical and linear selectivity in human language processing. We replicate

this experiment, to the extent possible,[1] on a series of large pretrained language models (§2). We design a series of superficially similar but structurally distinct grammaticality judgment tasks. We generate hierarchical grammars that accord to natural language structure, as well as linear grammars that are explained by positional insertion or transformation rules. Using these models and stimuli, we investigate the following research questions:

RQ1. Do models process hierarchically-structured inputs in a distinct manner from linearly-structured inputs? *We find that LLMs demonstrate distinct behaviors and mechanisms for grammars defined by hierarchical versus linear structure.*

RQ2. Which model components judge grammaticality in hierarchical vs. linear inputs, and to what extent are they shared? *We find strong overlap among hierarchical grammars and low overlaps between hierarchical and linear ones. Removing hierarchy-sensitive components significantly reduces accuracy for hierarchical grammars while having a smaller effect on linear grammars.*

RQ3. Are the findings from **RQ1** and **RQ2** dependent on grounding in the lexicon of the language(s) of the training corpus? Or do these distinctions also hold given grammars generated using nonce words? *We observe that the natural-language hierarchy-sensitive components also have significant influence on nonce grammars, suggesting that these components are more selective for hierarchical structure than in-distribution language.*

These results provide evidence that model regions responsible for processing hierarchical linguistic structure are localizable and distinct. Further, these regions are selective for hierarchically structured language more broadly, and are not dependent on meaning nor in-distribution language inputs. This suggests that functional specialization toward hierarchical linguistic structure can arise solely from exposure to language data. Thus, even in the absence of strong human-like inductive biases, language-selective regions can emerge.[2]

[1] Musso et al.'s (2003) experiment required that subjects be fluent in their native language (German, in their case) and not have prior exposure to the foreign languages (Italian and Japanese). LLMs' training distributions contain documents in non-English languages—though orders-of-magnitude fewer documents than for English. Thus, while we cannot fully satisfy this condition for LLMs, we can distinguish dominant and non-dominant languages of the model's training corpus.

[2] Data and code are available at this anonymous Zenodo repository. This will be replaced with a GitHub link in the final version.

2 Methods

2.1 Models

We use Mistral-v0.3 (7B; Jiang et al., 2023), QWen 2 (0.5B and 1.5B; Yang et al., 2024), Llama 2 (7B; Touvron et al., 2023), and Llama 3.1 (8B and 70B; Grattafiori et al., 2024). We select these models because they are open-weights, relatively commonly used, and are currently among the best-performing open models. In all experiments, we use nucleus sampling (temperature = 0.1, p = 0.9). We run experiments on a node with 4 A100s (80G).

2.2 Data

We define 3 classes of hierarchical and linear grammars respectively in English, Italian, and Japanese, yielding 18 grammars total. The choice of languages aligns with the original Musso et al. study, which examined German, Italian, and Japanese. In Musso et al., the human participants' native language was German; as the LLMs we use are primarily trained on English, we substitute German with English while retaining Italian and Japanese. These represent a minimal sample of non-English languages that are typologically similar (Italian) and typologically distinct (Japanese) from the primary language of the LLMs we investigate.

Hierarchical and linear grammars differ in whether the grammaticality is explained by a hierarchical or positional rule. Hierarchical grammars contain rules that conform to the hierarchical structure of natural language (Chomsky, 1957; Everaert et al., 2015) while linear grammars, argued to be impossible in human language (Chomsky, 1957, 1965), contain rules that are defined by word positions or relative word orderings—e.g., insert a word at position 4.

We generate positive and negative examples for all grammars. A **positive** example follows the grammar rule, while a **negative** example violates it.[3] For hierarchical grammars, negative examples are created by swapping the final two words of a positive example. For linear grammars involving word insertion, negative examples result from inserting the word at the final position. For linear grammars that invert word order, negative examples are formed by swapping the final two words after reversing the input. For the linear grammar Italian last-noun agreement, a positive example

[3] Negative examples do *not* aim to convert hierarchical inputs into linear ones or vice versa; they simply violate the defined grammar rule.

	Grammar	Positive Example	Negative Example
Hierarchical	**Declarative**. Subject, verb, object.	a woman reads a chapter	a woman reads chapter a
	Subordinate. Subject, verb taking a relative clause complement.	Sheela thinks that the woman reads the chapter	Sheela thinks that the woman reads chapter the
	Passive. Like **Declarative**, but in the passive voice.	a chapter is read by a woman	a chapter is read by woman a
Linear	**Negation**. Insert "doesn't" or "don't" at position 5.	a woman reads a doesn't chapter 1 2 3 4 5 6	a woman reads a chapter doesn't 1 2 3 4 5 6
	Inversion. Invert the word order of **Declarative**.	chapter a reads woman a 5 4 3 2 1	chapter a reads a woman 5 4 3 1 2
	Wh-word. Insert wh-word at position 6.	did a woman reads a when chapter? 1 2 3 4 5 6 7	did a woman reads a chapter when? 1 2 3 4 5 6 7

Table 1: **Dataset.** Hierarchical and linear grammars, descriptions of the rule defining each grammar, and positive (grammatical) and negative (ungrammatical) examples for each. We provide only English examples here for space; see App. A.1 for descriptions and examples for all grammars. Also see § 2.2 for detailed descriptions of hierarchical and linear grammars, and the details of positive and negative example construction.

aligns the determiner's gender with the last noun, while a negative example incorrectly matches the determiner to the first noun. Sentences are generated using templates based on Musso et al. (2003). Each grammar, its rule, and examples are provided in Tables 1 and 2. Our dataset includes 7 verbs with at least 5 subjects and objects each, yielding 1106 positive-negative pairs per grammar. We use a 50/50 train-test split (553 pairs per grammar).

3 Experiments

We conduct four experiments to evaluate the behaviors and mechanisms of six LLMs when processing hierarchical and linear grammars in an in-context learning setup. These models are pre-trained primarily on English but also include substantial amounts of other high-resource languages. While the exact training composition remains unknown, LLMs are typically trained on web-scraped data,[4] suggesting a predominance of English text (W3Techs, 2024), alongside significant content from other widely used languages.

Experiment 1 evaluates pre-trained LLMs on grammaticality judgment tasks for hierarchical and linear grammars (§3.1). Experiment 2 identifies model components crucial for processing these structures by treating hierarchical and linear inputs as counterfactuals (§3.2). Experiment 3 tests the causal role of these components by ablating them and measuring changes in grammaticality judgments (§3.3). Experiment 4 examines whether these components reflect mere in-distribution generalization or a broader sensitivity to hierarchical and linear structure using nonce sentences (§3.4).

[4] Relevant discussion on Huggingface.

The input prompt in our in-context learning setup comprises ten demonstrations, followed by a test example (details in §3.1). In all experiments, we conduct four trials, presenting mean results across four random seeds; demonstrations in the input prompt are randomized between trials, while test examples remain consistent. The format of the prompts remain consistent across experiments.

3.1 Experiment 1: Are language models significantly more accurate at classifying the grammaticality of sentences from hierarchical grammars?

We evaluate LLM accuracy on grammaticality judgments across different grammars. Musso et al. (2003) found that humans classify hierarchical grammars more accurately, even without prior fluency in the test languages. If LLMs contain functionally specialized regions for hierarchical processing, we expect a similar pattern. As described in §2, we generate 1106 examples for each of the 18 grammars and perform a uniform 50/50 train/test split. Each LLM is prompted with an instruction describing the task (see Appendix B.1.1), followed by 10 in-context demonstrations sampled from the training set. These demonstrations use the format "`Q: {sentence}\nA: {answer}`", where `answer` is Yes for positive and No for negative examples. Each prompt contains exactly 5 positive and 5 negative examples in random order. The model then performs a metalinguistic judgment task, generating " Yes" or " No" for test examples.[5] We extract probabilities for " Yes" and " No" to assess correctness, reporting accuracy in Figure 1.

[5] The leading space in the answer tokens is intentional, matching the expected tokenization.

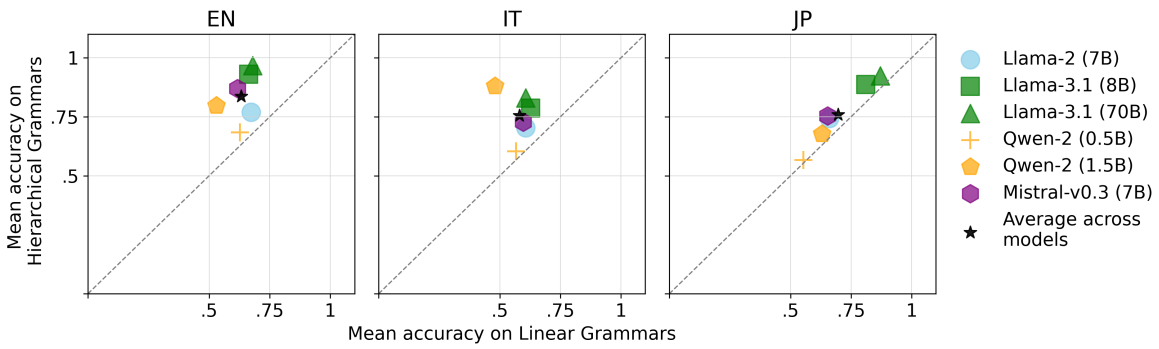


Figure 1: Few-shot accuracy on the grammaticality judgment task on hierarchical and linear inputs. On average, all models are better at the grammaticality judgment task on hierarchical inputs as compared to linear inputs. On hierarchical grammars, models are best at processing English inputs followed by Italian and Japanese. Model-wise accuracy on this task is shown in Figure 5 in App. B.1. Grammar-wise accuracy is shown in Table 5 in App. B.1.

Hypothesis. Natural language is largely ambiguous with respect to linear versus hierarchical structure (Chomsky, 1957); human brains have biological preferences for hierarchical structures (Musso et al., 2003), but LLMs do not have this preference built into their architecture (Min et al., 2020; McCoy et al., 2018; Mueller et al., 2022), so it is not clear *a priori* whether they would treat these structures in the same way. Given results from Kallini et al. (2024), we hypothesize that models will be significantly more accurate when labeling sentences from hierarchical than linear grammars. We also expect larger models to be more accurate.

Results. We find (Figure 1) that for English and Italian grammars, models are better at distinguishing positive and negative examples in hierarchical grammars than linear grammars ($p < .001$; see Table 4). This difference is greater for larger models than smaller ones, perhaps indicating greater functional specialization with scale. This provides initial support for our hypothesis that hierarchical and linear grammars are processed in distinct manners.

3.2 Experiment 2: Are the model components implicated in processing hierarchical structures disjoint from those implicated in processing linear structures?

Our behavioral evaluations suggest that LLMs are more accurate on grammaticality judgment tasks with hierarchical inputs, but this does not disambiguate whether models have separate *mechanisms*[6] for processing hierarchical and linear grammars. If a model has specialized mechanisms for processing hierarchical and linear grammars, we hypothesize that the set of model components causally responsible for correct grammaticality predictions on hierarchical inputs should be different from those responsible for correct predictions on linear inputs.

To test this, we locate neurons in the model that are most sensitive towards processing hierarchical syntax. Specifically, we investigate dimensions of the output vector of the MLP and attention submodules in each layer.[7] We test whether there is significant overlap between the neurons responsible for processing hierarchical and linear structures.

Recall that we prompt the model with a task instruction followed by 10 uniformly sampled demonstrations of positive and negative examples. Given this prompt, we quantify the importance of each neuron z in increasing the logit difference m between the correct and incorrect answer tokens y and y' for a test sentence t. In other words, given a language model $\mathcal{M}$, $m = \mathcal{M}(t)_{y'} - \mathcal{M}(t)_y$; $\mathcal{M}(t)_y$ and $\mathcal{M}(t)_{y'}$ are the logits corresponding to the correct and incorrect answer tokens. We compute the component z's **indirect effect** (IE; Pearl, 2001; Robins and Greenland, 1992) on m given the test sentence t, and a minimally different sentence t' that flips the correct answer from y to y'.[8] Acti-

[6] We use "mechanism" to refer to a causal chain proceeding from an initial cause to a final effect. In language models, this refers to a set of causally implicated model components that explain how inputs are transformed into the observed output behavior m, which we define below.

[7] For MLPs, we use the output of the down-projection *after* the non-linear transformation. For attention, we use the output of the out projection.

[8] If t is a positive example, then t' is the corresponding negative example formed by swapping the appropriate word(s) or modifying the sentence. If t is a negative example, then t' is the corresponding positive example.

vation patching (Vig et al., 2020; Finlayson et al., 2021; Geiger et al., 2020; Meng et al., 2022), a common procedure for computing the IE of model components, entails computing the IE as follows:

$$\text{IE}(m; z; t, t') = m(t|\text{do}(z_t = z_{t'})) - m(t) \quad (1)$$

Activation patching is computationally expensive, as the number of required forward passes scales linearly with the number of neurons. Therefore, we instead use attribution patching (Kramár et al., 2024; Syed et al., 2024), a first-order Taylor approximation of the IE:

$$\hat{\text{IE}}(m; z; t, t') = \nabla_z m|_t \, (z_{t'} - z_t) \quad (2)$$

$\hat{\text{IE}}$ can be computed for *all* z using only 2 forward passes and 1 backward pass; i.e., the number of passes is constant with respect to the number of neurons. While not a perfect approximation, $\hat{\text{IE}}$ correlates almost perfectly with IE in typical cases (Kramár et al., 2024; Marks et al., 2024).[9]

We select the top 1% of both attention and MLP neurons in the model by $\hat{\text{IE}}$. We compute the pairwise overlap of this top 1% neuron subset for each pair of grammars to measure mechanistic overlap.

Hypothesis. If there are distinct mechanisms for processing hierarchical and linear grammars, there should be significant overlap between pairs of hierarchical structures, and significant overlap between pairs of linear structures. However, overlaps *across* linear and hierarchical structures should be significantly lower than overlaps between pairs of hierarchical grammars or pairs of linear grammars.

Results. We first observe that all mean pairwise component overlaps are significantly different from 0 (Figure 2). However, this overlap is significantly higher ($p < 0.001$) between pairs of hierarchical grammars than across pairs of hierarchical and linear grammars (See Table 8 in App. B.2 and Figure 2). This holds across English, Italian, and Japanese. This supports the hypothesis that LLMs use specialized components for processing hierarchical syntax that are distinct from those responsible for processing linear syntax.

We also observe that linear structures that share a rule across languages, such as inversions, show stronger overlaps than arbitrary pairs of linear structures (Figures 8 and 11 in Appendix B.2). This serves as a sanity check that the component overlaps correlate with structural similarities.

[9]Except at the first and last layer, where the correlation is still strong but significantly lower.

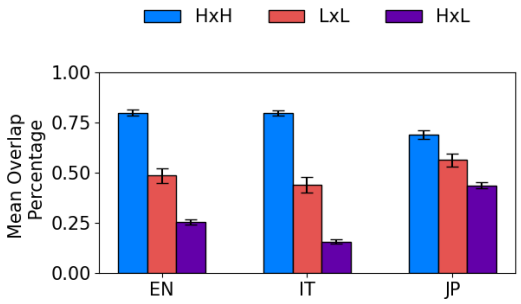


Figure 2: Mean pairwise overlap percentage of the top 1% of neurons from hierarchical (H) or linear (L) grammars. We show means across models (error bars are standard errors); see Figure 7a in App. B.2 for model-wise results. Overlaps are significantly ($p < 0.001$, Table 8) different between hierarchical-hierarchical pairs and linear-linear pairs, and between hierarchical-hierarchical pairs and hierarchical-linear pairs.

3.3 Experiment 3: Does ablating hierarchy-sensitive components affect performance on linear grammars, and vice versa?

We have located neurons responsible for processing hierarchical and linear grammars. If these neurons are selective for only hierarchical or linear structure, then ablating them should selectively impact the model's performance on the grammaticality judgment tasks from §3.1. We now perform an ablation experiment to causally verify this prediction.

Let $\bar{a}_i$ be the mean activation of neuron a at token position i across training examples. We first cache $\bar{a}_i$ for each MLP and attention output dimension. We then run three additional iterations of the grammaticality judgment task from §3.1, each while ablating a different set of components. (i) We ablate the union of the top 1% of neurons by $\hat{\text{IE}}$ across hierarchical grammars. (ii) We take the union of the top 1% of neurons across linear grammars, subsample to the same number of neurons as in the hierarchical union (subsampling procedure described below), and ablate this set. Finally, (iii) we ablate a random uniform subsample of neurons, where the number of ablated neurons is the same as in (i) and (ii). Sets (i) and (ii) are derived from §3.2. We call the hierarchy-sensitive neuron set H and the linearity-sensitive neuron set L.

Due to the strong overlaps between components responsible for processing hierarchical syntax and only minimal overlaps between components responsible for processing linear syntax, we observe

that $|L| \approx 2|H|$. We therefore subsample L to be the same size as H by (1) sorting components in L by their effect size (as found in §3.2) and (2) keeping the top $|H_\ell|$ components from layer ℓ, where $|H_\ell|$ is the size of H at layer ℓ. When ablating the uniform subsample, we uniformly sample and ablate $|H_\ell|$ components in each layer ℓ.

Hypothesis. If H and L are functionally distinct, then ablating H should reduce performance on hierarchical grammars more than ablating L and more than ablating random components. Ablating L should reduce performance on linear grammars more than ablating H and more than ablating random components.

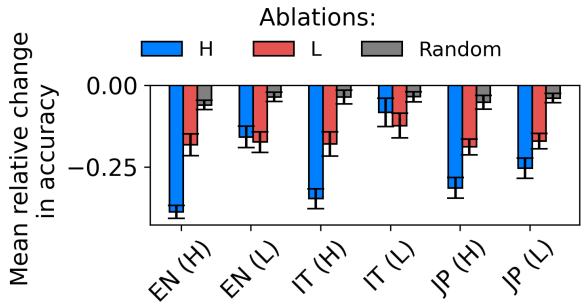


Figure 3: Mean relative change in accuracy across models (error bars are standard errors) after ablating the top 1% of neurons from hierarchical (H) or linear (L) grammars. We compare to a random ablation baseline. For model-wise ablations, see Figure 12a in App. B.3.

Results. Ablating components from H decreases the models' accuracy on hierarchical structures significantly more than ablating components in L (Figure 3; see Table 9 in App. 3.3 for significance tests). Ablating components from L decreases the model's accuracy on linear structures more than ablating H. Ablating uniformly sampled components causes a lower decrease in performance compared to ablations from the H or L sets.

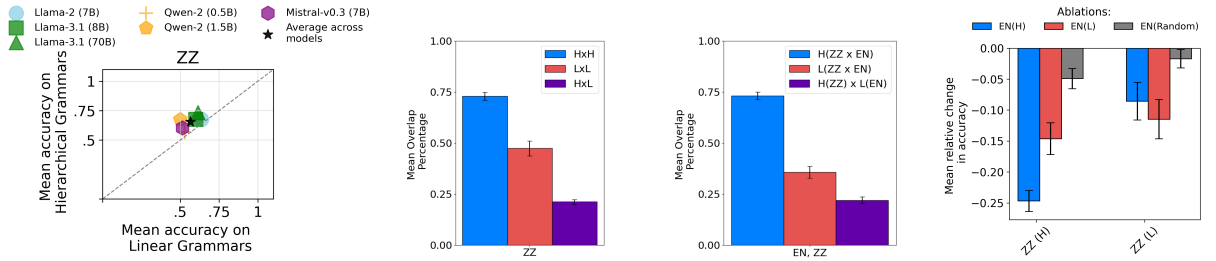
These results are further mediated by model and language. Llama-3.1, Mistral-v0.3, and QWen-2 (1.5B) show larger decreases in relative accuracy on hierarchical and linear inputs when ablating the H and L sets, respectively. Llama-2 and QWen-2 (0.5B) show similar changes in performance under ablations, though not in the selective manner we observe for other models. At the language level, these trends are consistent across English and Italian, but only sometimes generalize to Japanese. Overall, our results suggest that for Llama-3.1, Mistral-v0.3, and QWen (1.5B), the components discovered in §3.2 selectively reduce model performance in an expected manner in English and Italian. For other models, there is more mechanistic overlap in how grammatically judgments are performed for hierarchical and linear inputs. Thus, we largely find support for the hypothesis of hierarchical functional selectivity. Exceptions include smaller models, and results in Japanese (a less frequent language in the training corpora of these models); this provides preliminary evidence that greater functional specialization may emerge with scale, both with respect to dataset size and number of parameters.

3.4 Experiment 4: Are these neurons sensitive to hierarchical structure or in-distribution language?

Thus far, our results have been confounded by the fact that hierarchical sentences are commonly attested in the natural language corpora that LLMs are trained on, whereas linear sentences would be unlikely to appear. Thus, it is unclear if we have observed functional selectivity toward hierarchical language, or merely toward *in-distribution* language. To address this confound, we propose additional experiments using sentences constructed from nonce words—what Fedorenko et al. (2016) call **Jabberwocky sentences** (abbreviated ZZ).

We define a bijective mapping from all words in the English grammars to nonce words; see Table 3 for examples. Then, we replicate our previous experiments on this set of out-of-distribution Jabberwocky sentences. By preserving the distinction between hierarchical and linear grammars and using a meaningless lexicon, we can disentangle hierarchy-sensitive mechanisms from mechanisms that are merely sensitive to natural language distributions resembling those in the training corpus.

Hypotheses. In humans, Jabberwocky sentences cause a smaller increase in neural activity as compared to natural sentences (Fedorenko et al., 2016), implying that the language processing regions of the brain are *not* sensitive to Jabberwocky sentences. If language models are similarly selective for meaningful inputs—and therefore, if the H neurons from previous experiments are actually in-distribution-language neurons, and if the L neurons are actually out-of-distribution language neurons—then we expect the following three trends. (1) There should not be significant differences in model performance on grammaticality judgments



(a) (b) (c) (d)

Figure 4: Results on Jabberwocky grammars. We show grammaticality judgment task performance (a), mean neuron overlap percentages between Jabberwocky hierarchical and linear grammars (b), neuron overlaps between English and Jabberwocky grammars (c), and the mean relative changes in accuracy as measured on Jabberwocky grammars after ablating top 1% of neurons corresponding to English grammars (d). See App. B.4 for model-wise results.

for hierarchical and linear Jabberwocky grammars. (2) There should be little overlap between the hierarchical English and Jabberwocky neurons; by implication, ablating neurons discovered from natural language inputs should not affect performance on Jabberwocky sentences. (3) It is not clear whether we should expect distinct mechanisms for processing H and L Jabberwocky grammars; if there is an abstract acceptability judgment circuit that is not tied to natural language, then it should be present in the linear natural-language neurons. Thus, we hypothesize that the L neurons from previous experiments will affect performance on Jabberwocky sentences more than the H neurons.

Results. In behavioral experiments using Jabberwocky sentences, we find (Figure 4a) that the gap in performance from hierarchical to linear grammars is significantly lower than that for English grammars—but still consistently present across models. The lower gap could be because performance is closer to chance than for natural grammars. The small gap in performance partially contradicts Hypothesis 1, but does not provide strong enough evidence to confidently reject it. Attribution patching results (Figure 4b) suggest that the components used to correctly judge hierarchical and linear Jabberwocky inputs are largely disjoint: overlaps between pairs of hierarchical structures are significantly higher than overlaps across pairs of hierarchical and linear grammars. Moreover, the components used to process hierarchical English grammars are strongly shared with the components that are used to process hierarchical Jabberwocky grammars (Figure 4c), while overlaps between linear English grammars and hierarchical Jabberwocky grammars is low. This contradicts Hypotheses 2 and 3, suggesting that the hierarchy-sensitive mechanisms we have observed in LLMs may be more abstract and generalized than those in humans. Thus, Jabberwocky and English grammars share LLM hierarchy-sensitive components, whereas humans show a smaller neural response to Jabberwocky compared to English sentences.

Lastly, we observe (Figure 4d) that ablating the top 1% of neurons from the English hierarchical grammars causes a significant decrease in accuracy when processing hierarchical Jabberwocky inputs; similar decreases in linear Jabberwocky accuracy result from ablating English linear components. This suggests that the causally relevant natural-language and Jabberwocky neuron sets are shared to a significant extent (see Table 11 in App. B.4). Further, ablations to English hierarchical components causes *selective* decreases in Jabberwocky hierarchical accuracy; selectivity is lower when ablating English linear components.

Taken together, these results provide evidence that LLMs' hierarchy-sensitive and linearity-sensitive component sets are sensitive primarily to the structure of the grammar, and only depend on grounding in meaning or in-distribution language to a minor extent. This provides significant ($p < .05$) and causal evidence against Hypothesis 2, which we reject. Results from Figure 4d and Table 11 also give sufficient evidence to reject Hypothesis 3. Thus, there exist grammaticality judgment mechanisms that are selective for hierarchical structure in a highly abstract manner, and that do not merely select for in-distribution language. That said, there are components are selective to both hierarchical *and* in-distribution language, but these do not make up the majority of the components found in previous experiments.

4 Discussion and Related Work

Acquiring syntax-selective subnetworks. We find behavioral and causal evidence supporting the hypothesis that hierarchical and linear grammars are processed using largely disjoint mechanisms in large language models. Thus, as in humans (Baker et al., 2007), general-purpose learners such as language models can *acquire* functionally specific regions. To some extent, linguistic functional selectivity in LLMs is surprising: humans process many more modalities and signal types than language alone, so functional specialization toward linguistic signals may be sensible as one among many modal specializations (Kanwisher, 2010). However, unimodal language models like those we test are exposed *only* to text. While not all of this text is natural language, one might expect a larger portion of the model to be responsible for processing hierarchical structure. These as well as our behavioral results extend prior evidence that pretraining induces preferential reliance on syntactic features over positional features (Mueller et al., 2022; Murty et al., 2023; Ahuja et al., 2024),[10] and supports prior findings that there exist syntax-selective—and more broadly, language-selective—subnetworks in LLMs (AlKhamissi et al., 2024; Sun et al., 2024).

Human-likeness and learnability. Note that hierarchical functional specialization is *not* evidence that humans and LLMs process language in the same manner. Fedorenko et al. (2016) find that language processing circuits in the brain activate significantly less on Jabberwocky sentences, whereas we observe significant overlaps (albeit not complete) in these circuits in LLMs. This suggests some degree of selectivity for natural in-distribution language, as in humans, but the hierarchy-sensitive mechanisms are also more abstract and not tied to meaning as in humans.

There is evidence that hierarchical grammars are easier to learn than grammars that do not occur in human languages (Kallini et al., 2024; Ahuja et al., 2024). This could provide an explanation for why language models are so attuned to this structure and learn to explicitly represent it: it is easier to learn a hierarchical organization than flat organization of a vocabulary, and it may simply be a more efficient explanation of the distribution. That said, randomly shuffling input data does not seem to destroy downstream performance (Sinha et al., 2021), despite destroying performance on structural probing tasks (Hewitt and Manning, 2019). Future work should investigate the relationship between the syntax-sensitive components we discover and performance on downstream NLP tasks.

Mechanistic interpretability. Using causal localizations to investigate the mechanisms underlying model behaviors has recently become more popular (e.g., Wang et al., 2023; Hanna et al., 2023; Prakash et al., 2024; Merullo et al., 2024; Bayazit et al., 2024). While localization is not equivalent to explanation, it can reveal distinctions in where and how certain phenomena are encoded in activation space. Future work could employ techniques from the training dynamics and mechanistic interpretability literature to better understand how and when these components arise during pretraining, as well as the (presumably numerous) functional sub-roles of these distinct component sets.

More broadly, this work suggests a less-explored direction in interpretability based on high-level coarse-grained abstractions. Much recent work has aimed to discover more fine-grained and single-purpose units of causal analysis (e.g., sparse autoencoder features; Bricken et al., 2023; Cunningham et al., 2024; Marks et al., 2024); we believe that a parallel direction based in functionally coherent sets (or subgraphs) of components would yield equally interesting insights. For example, effective representations of syntax are a necessary condition for robust language understanding and generation; thus, we would expect the hierarchy-sensitive components we discover to be implicated in *any* NLP task if the model were robustly understanding the inputs. Therefore, not relying on these components could be a signal that models have learned to rely on some mixture of heuristics.

5 Conclusion

We have investigated whether there exist localizable and functionally distinct sets of components for processing hierarchically versus linearly structured language inputs. We find behavioral and causal evidence that these component sets are distinct, both in location and in their functional role in the network.

[10] Note, however, that these behavioral results may be explainable using teleological approaches such as those in McCoy et al. (2024): linear grammaticality judgment is a low-probability task and contains low-probability inputs (assuming a pretraining distribution based on Internet text), and will therefore be more difficult for a language model to perform, even if the model used a shared mechanism to perform each grammaticality judgment task in this study.

Limitations

We acknowledge that our work could be improved in several respects. First, neurons and attention outputs are problematic units of analysis due to polysemanticity (Elhage et al., 2022); i.e., observing the activations of a component is often not informative, as they are sensitive to many features simultaneously. Further, the component sets we analyze are unordered sets, which means that we do not yet understand how many distinct mechanisms are responsible for the behaviors we observe, nor what these mechanisms qualitatively represent.

We have also not evaluated the effect of these components on tasks outside of grammaticality judgments. Thus, we do not yet understand how selective nor how robust these behaviors or localizations are under different settings. Further, the grammaticality judgment task may prime the model to be sensitive to valid linguistic structures more generally, rather than the structures that we present to the models; we therefore cannot confidently conclude that the significant accuracy differences we observe will generalize to other task settings or prompt formats given the same grammars.

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2024. Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2024. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. *arXiv preprint arXiv:2411.02280*.

Marie Amalric and Stanislas Dehaene. 2019. A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage*, 189:19–31.

Ian A Apperly, Dana Samson, Naomi Carroll, Shazia Hussain, and Glyn Humphreys. 2006. Intact first-and second-order false belief reasoning in a patient with severely impaired grammar. *Social neuroscience*, 1(3-4):334–348.

Chris I. Baker, Jia Liu, Lawrence L. Wald, Kenneth K. Kwong, Thomas Benner, and Nancy Kanwisher. 2007. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 104(21):9087–9092.

Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. 2024. Discovering knowledge-critical subnetworks in pretrained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6549–6583, Miami, Florida, USA. Association for Computational Linguistics.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Paul Broca. 1861. Remarques sur le siége de la faculté langage articulé; suives d'une observation d'aphémie. *Bulletins et mémoires de la Société Anatomique de Paris*, 6:330–357.

Xuanyi Chen, Josef Affourtit, Rachel Ryskin, Tamar I Regev, Samuel Norman-Haignere, Olessia Jouravlev, Saima Malik-Moraleda, Hope Kean, Rosemary Varley, and Evelina Fedorenko. 2023. The human language system, including its inferior frontal component in "broca's area," does not support music perception. *Cerebral Cortex*, 33(12):7904–7929.

Noam Chomsky. 1957. *Syntactic structures*. De Gruyter Mouton.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. The MIT Press.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Martin BH Everaert, Marinus AC Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

Evelina Fedorenko and Rosemary Varley. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153.

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.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc

Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna A Ivanova, Zachary Mineroff, Vitor Zimmerer, Nancy Kanwisher, Rosemary Varley, and Evelina Fedorenko. 2021. The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, 2(2):176–201.

Anna A Ivanova, Shashank Srikant, Yotaro Sueoka, Hope H Kean, Riva Dhamala, Una-May O'reilly, Marina U Bers, and Evelina Fedorenko. 2020. Comprehension of computer code relies primarily on domain-general executive brain regions. *elife*, 9:e58906.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Nancy Kanwisher. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*.

Y Liu, J Kim, C Wilson, and M Bedny. 2020. Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. biorxiv, 2020.05. 24.096180.

Saima Malik-Moraleda, Maya Taliaferro, Steve Shannon, Niharika Jhingan, Sara Swords, David J Peterson, Paul Frommer, Marc Okrand, Jessie Sams, Ramsey Cardwell, et al. 2023. Constructed languages are processed by the same brain mechanisms as natural languages. *bioRxiv*.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018*, Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, pages 2096–2101. The Cognitive Science Society.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Martin M Monti, Lawrence M Parsons, and Daniel N Osherson. 2009. The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences*, 106(30):12554–12559.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. 2003. Broca's area and the language instinct. *Nature neuroscience*, 6(7):774–781.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

James M. Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haiyang Sun, Lin Zhao, Zihao Wu, Xiaohui Gao, Yutao Hu, Mengfei Zuo, Wei Zhang, Junwei Han, Tianming Liu, and Xintao Hu. 2024. Brain-like functional organization within large language models. *arXiv preprint arXiv:2410.19542*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *The 7th BlackboxNLP Workshop*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rosemary Varley and Michael Siegal. 2000. Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, 10(12):723–726.

Rosemary A Varley, Nicolai JC Klessinger, Charles AJ Romanowski, and Michael Siegal. 2005. Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9):3519–3524.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Jesse Vig, 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.

W3Techs. 2024. Usage statistics and market share of content languages for websites, may 2024. Accessed: 2024-05-18.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

A Methods

A.1 Grammar rule descriptions

We define a series of hierarchical sentences in English, Japanese, and Italian.

- **Declarative sentence**: For English sentences, subjects and objects can be singular or plural nouns. Verbs agree with their subjects. IT sentences are Italian translations of the English sentences. Unlike Italian and English which have SVO word order, Japanese translations (JP sentences) have SOV word order.

- **Subordinate sentence**: In each language, matrix subjects, subordinate subjects, matrix objects, and subordinate objects can be singular or plural nouns. In English and Italian, verbs of the subordinate subject and the subject agree with their respective subjects in number. We generate subordinate clauses by using verbs which take complementizer phrases as objects (e.g., "Tom sees that the dog carries the fish"). English and Italian both place the main clause's verb before the start of the subordinate clause, whereas Japanese places the main verb after the end of the clause.

- **Passive sentence**: Subjects and objects can be singular or plural nouns. Verbs are always in the passive form. Like in (Musso et al., 2003), in the passive construction, we include the agent of a transitive verb in a prepositional phrase.

- **Null subject sentence**: This structure is restricted to Italian. We use the verb and object without the subject, since the use of the subject is not a strict requirement in Italian.[11]

Linear Grammars Similar to Musso et al. (2003), the linear sentences we test are constructed by breaking the hierarchical order between the subject and the nominal words. While our linear sentences use English, Italian, and Japanese lexicons, they break the hierarchical relationship between

[11] Italian verbal morphology provides all person and number information needed to understand the subject of a sentence, whereas English morphology does not provide this information. That said, there exist languages without the verbal person/number inflection that optionally allow dropping the subject of the sentence if it is the topic of that sentence, such as Mandarin and Japanese; thus, this structure is still attested and therefore still qualifies as a hierarchical (UG-compliant) structure.

the subject, verb, and object, using the strategies described below.

- **Negation**: We break the hierarchical order by inserting a negation word "doesnt́" after the fifth word in English sentences. In Italian, we insert 'non' (IT) after the third word. In Japanese, we insert ない (JP) after the third word.
- **Inversion**: We invert the order of the words in a sentence (before tokenization) to form the second construction.

The third construction varies between languages.

- **Wh-word** (English): We include a question in the subordinate clause of the sentence by inserting a 'wh-' word (who, why, what etc.) at the penultimate token position.
- **Last noun agreement** (Italian): We change the subject term's gender to always match that of the final noun in the sentence.
- **Past Tense** (Japanese): The Japanese past tense construction was built by adding the suffix -ta, not on the verb element as in the hierarchical grammatical rule for Japanese, but on the third word, counting from right to left.

A.2 Dataset Description and Examples

Examples of all the grammars we construct in English, Italian, and Japanese may be found in Table 2. Examples of Jabberwocky sentences may be found in Table 3.

B Experiments

B.1 Experiment 1: Few-shot learning accuracy

Experiment 1 (§3.1) assesses the model's performance on grammaticality judgments of hierarchical and linear structures. Here we share statistical comparisons of the accuracy distributions (Table 4), accuracy values by language (Figure 5 and Table 5) and grammar-wise accuracy values (Table 5).

B.1.1 Example Prompts

We present example prompts from one of the hierarchical structures for each language. The prompt skeleton is in English, regardless of the language used for the examples. We intentionally strip the final whitespace after `A:`, as the model expects a leading space within the answer token (and thus, it should not be present in the prompt).

English example.

```
"Here are English sentences that either follow or break a grammar rule. Each sentence is labeled 'Yes' if it follows the rule and 'No' if it doesn't. Label the final sentence as 'Yes' or 'No' based on whether it follows the same rule.

Q: Is this sentence grammatical? Yes or No: a woman drinks espresso the
A: No

Q: Is this sentence grammatical? Yes or No: the architects touch a mouse
A: Yes

Q: Is this sentence grammatical? Yes or No: the women eat cucumber the
A: No

Q: Is this sentence grammatical? Yes or No: the writers drink a lemonade
A: Yes

Q: Is this sentence grammatical? Yes or No: a teacher touches a lightbulb
A: Yes

Q: Is this sentence grammatical? Yes or No: the actress touches toy a
A: No

Q: Is this sentence grammatical? Yes or No: a boy kicks bottle a
A: No

Q: Is this sentence grammatical? Yes or No: the woman pushes toy a
A: No

Q: Is this sentence grammatical? Yes or No: a professor reads a poem
A: Yes

Q: Is this sentence grammatical? Yes or No: the orators read a story
A: Yes

Q: Is this sentence grammatical? Yes or No: the doctor drinks milkshake the
A:"
```

	Language	Grammar	Positive Example	Negative Example
Hierarchical	English (EN)	**Declarative**	a woman reads a chapter	a woman reads chapter a
		Subordinate	Sheela thinks that the woman reads the chapter	Sheela thinks that the woman reads chapter the
		Passive	a chapter is read by a woman	a chapter is read by woman a
	Italian (IT)	**Declarative**	una donna legge un capitolo	una donna legge capitolo un
		Subordinate	Sheela pensa che una donna legge un capitolo	Sheela pensa che la donna legge capitolo un
		Passive	un capitolo è letto da una donna	un capitolo è letto da donna una
	Japanese (JP)	**Declarative**	女性 は 章 を 読む	女性 は 章 読む を
		Subordinate	シーラ は 女性 が 章 を 読む と 考える	シーラ は 女性 が 章 を 読む 考える と
		Passive	章 は 女性 に 読まれる	章 は 女性 読まれる に
Linear	English (EN)	**Negation**. Insert "doesn't" or "don't" at position 5.	a woman reads a doesn't chapter 1 2 3 4 5 6	a woman reads a chapter doesn't 1 2 3 4 5 6
		Inversion. Invert the declarative word order.	chapter a reads woman a 5 4 3 2 1	chapter a reads a woman 5 4 3 1 2
		Wh-word. Insert wh-word at position 5.	did a woman reads a when chapter? 1 2 3 4 5 6 7	did a woman reads a chapter when? 1 2 3 4 5 6 7
	Italian (IT)	**Negation**. Insert "no" at position 5.	una donna legge un no capitolo 1 2 3 4 5 6	una donna legge un capitolo no 1 2 3 4 5 6
		Inversion. Invert the declarative word order.	capitolo un legge donna una 5 4 3 2 1	capitolo un legge una donna 5 4 3 1 2
		Last-noun agreement. Make all determiners agree with the gender of the final noun.	~~una~~ un donna legge un capitolo 1 2 3 4 5	una donna legge un capitolo 1 2 3 4 5
	Japanese (JP)	**Negation**. Insert a negation word at position 4.	女性 は 章 ない を 読む 1 2 3 4 5 6	女性 は 章 を 読む ない 1 2 3 4 5 6
		Inversion. Invert the declarative word order.	読む を 章 は 女性 5 4 3 2 1	読む を 章 女性 は 5 4 3 1 2
		Past tense. Insert the past tense marker at position 4.	女性 は 章 をた 読む 1 2 3 4 5	女性 は 章 読む をた 1 2 3 4 5

Table 2: **Dataset.** List of grammars, descriptions of the rule defining each grammar, and corresponding positive and negative examples.

Italian example. "Here are Italian sentences that either follow or break a grammar rule. Each sentence is labeled 'Yes' if it follows the rule and 'No' if it doesn't. Label the final sentence as 'Yes' or 'No' based on whether it follows the same rule.

Q: Is this sentence grammatical? Yes or No: una donna beve espresso il
A: No

Q: Is this sentence grammatical? Yes or No: l' architette toccano il topo
A: Yes

Q: Is this sentence grammatical? Yes or No: le donne mangiano cetriolo il
A: No

Q: Is this sentence grammatical? Yes or No: le scrittrici bevono la limonata
A: Yes

Q: Is this sentence grammatical? Yes or No: un' insegnante tocca una lampadina

	Grammar	Positive Example	Negative Example
Hierarchical	**Declarative**. Subject, verb, object.	a wug ungos the snorfle	a wug ungos snorfle the
	Subordinate. Subject, verb taking a relative clause complement.	Gomu herdles that the blincos stoffle the pelunko	Gomu herdles that the blincos stoffle reads pelunko the
	Passive. Like **Declarative**, but in the passive voice.	the gunzle is snugoed by the wugen	the gunzle is snugoed by wugen the
Linear	**Negation**. Insert "doesn't" or "don't" at position 5.	the arcuplos ungo a doesn't blorft 1 2 3 4 5 6	the arcuplos ungo a blorft doesn't 1 2 3 4 5 6
	Inversion. Invert the word order of **Declarative**.	snorfle the ungos wug a 5 4 3 2 1	snorfle the ungos a wug 5 4 3 1 2
	Wh-word. Insert wh-word at position 5.	Did a knurkle gurdles a when skerpo? 1 2 3 4 5 6 7	Did a knurkle gurdles a skerpo when? 1 2 3 4 5 6 7

Table 3: **Jabberwocky dataset.** List of grammars, descriptions of the rules defining each grammar, and positive (grammatical) and negative (ungrammatical) examples for each. We use similar prompt constructions as in the English examples (Also see §B.1.1).

```
A: Yes

Q: Is this sentence grammatical? Yes or
No: l attrice tocca giocattolo un
A: No

Q: Is this sentence grammatical? Yes or
No: un ragazzo calcia bottiglia una
A: No

Q: Is this sentence grammatical? Yes or
No: la donna spinge giocattolo un
A: No

Q: Is this sentence grammatical? Yes or
No: una professoressa legge un poema
A: Yes

Q: Is this sentence grammatical? Yes or
No: gli oratori leggono la storia
A: Yes

Q: Is this sentence grammatical? Yes or
No: la dottoressa beve frappè il
A:"
```

Japanese example. "Here are Japanese sentences that either follow or break a grammar rule. Each sentence is labeled 'Yes' if it follows the rule and 'No' if it doesn't. Label the final sentence as 'Yes' or 'No' based on whether it follows the same rule.

```
Q: Is this sentence grammatical? Yes or
No: 女性はエスプレッソ飲むを
A: No

Q: Is this sentence grammatical? Yes or
No: 建築家たちはマウスを触る
A: Yes

Q: Is this sentence grammatical? Yes or
No: 女性たちは胡瓜食べるを
A: No

Q: Is this sentence grammatical? Yes or
No: 作家たちはレモネードを飲む
A: Yes

Q: Is this sentence grammatical? Yes or
No: 教師は電球を触る
A: Yes

Q: Is this sentence grammatical? Yes or
No: 女優は玩具触るを
A: No

Q: Is this sentence grammatical? Yes or
No: 少年はボトル蹴るを
A: No

Q: Is this sentence grammatical? Yes or
No: 女性は玩具押すを
A: No

Q: Is this sentence grammatical? Yes or
No: 教授は詩を読む
A: Yes

Q: Is this sentence grammatical? Yes or
No: 演説家たちは小説を読む
A: Yes
```

Q: Is this sentence grammatical? Yes or No: 医者はミルクセーキ飲むを
A:"

B.2 Experiment 2: Identify MLP and Attention Components with the highest $\hat{IE}$

Experiment 2 locates MLP and Attention neurons that are implicated in processing hierarchical and linear structures, and investigates if these components are disjoint. The language-level pairwise overlaps for all models across English, Italian and Japanese hierarchical and linear inputs, as well as Jabberwocky hierarchical and linear inputs is given in Figures 7. Grammar wise overlaps for MLP and attention components for English, Italian, and Japanese grammars are shown in Figures 8 and 11, respectively. Grammar wise overlaps for MLP and attention components for Jabberwocky grammars are shown in Figures 8 and 11, respectively. We also share results testing whether the pairwise mean overlaps of H, L, and HxL structures are significantly different among grammars using English, Italian, and Japanese versus Jabberwocky tokens in Tables 8 and 7.

B.3 Experiment 3: Ablations of top 1% of Attention and MLP Components

Experiment 3 considers selective ablations of hierarchy and linearity sensitive components, and evaluates how these ablations impact the accuracy of the model on the in-context learning task. We share ablation results by model in Figure 12a for English, Italian, and Japanese grammars. Through model-wise comparisons, we find that relative accuracy decreases on hierarchical grammars are significantly different for English, Italian and Japanese grammars depending on whether hierarchical, linear, or uniformly sampled components are ablated (see Table 9). However, the same is not true for linear grammars—relative accuracy decreases are not significantly different between ablations of hierarchical/linear components. In the case of Italian structures, ablating hierarchy-sensitive components appears akin to ablating uniformly sampled components.

Additionally, we also run ablation experiments on jabberwocky grammars using components that are sensitive to hierarchical or linear jabberwocky grammars. We share ablation results by model for jabberwocky grammars in Figure 12b. Here also we find that ablating hierarchy versus linearity sensitive components can cause a significant difference in the decrease in accuracy on jabberwocky hierarchical grammars relative to the no ablation case. This is not true for jabberwocky linear grammars (See Table 10).

B.4 Experiment 4: Are neurons identified in experiment 3 sensitive to hierarchical structure or in-distribution lexical tokens?

Experiment 4 considers selective ablations of the top 1% of hierarchy and linearity sensitive components, and evaluates how these ablations impact the accuracy of the model on the in-context learning task, when processing Jabberwocky grammars. If the neurons discovered in Experiment 3 are *not* sensitive to hierarchical structure and instead sensitive to in-distribution tokens, these ablations should not cause a decrease in model performance on Jabberwocky grammars which are composed of meaningless words. Alternatively, any decreases in model performance on Jabberwocky grammars should be caused by neurons in the L set which are, say, sensitive to out of distribution inputs. Ablation results by model are in Figure 13. We also present grammar-wise overlaps of the top 1% of attention and MLP neurons for hierarchical and linear English and Jabberwocky grammars in Figures 15 and 14 respectively, and show that the difference in overlaps between these grammars is statistically significant in Table 11. Then, we test the relative change in accuracy in Jabberwocky grammars after ablating components sensitive to hierarchical and linear English structures as well as uniformly sampled components. Ablating hierarchy vs. linearity sensitive components that are sensitive to the English task, causes a significantly different decrease in model performance on Jabberwocky hierarchical grammars. However, the same is not true for linear Jabberwocky grammars where ablating hierarchy sensitive components is no different from ablating linearity-sensitive or uniformly sampled components (see Table 12. This suggests that the components identified in Experiment 3 are at least partially sensitive to the structure of the inputs.

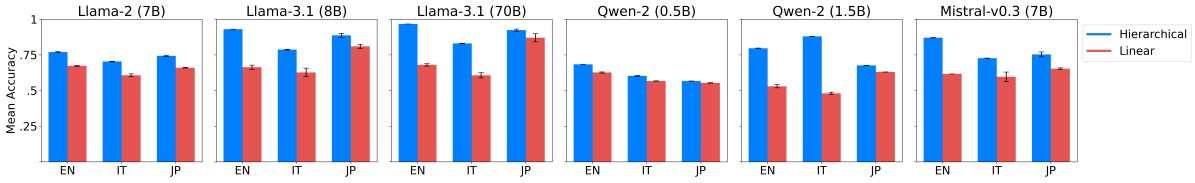


Figure 5: **Experiment 1.** Model-wise accuracy on the grammaticality judgments task given hierarchical and linear inputs from English, Italian and Japanese(See § 3.1 and Tables 5 and 2)

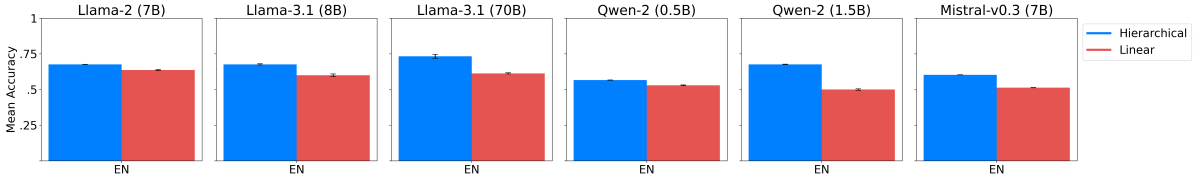
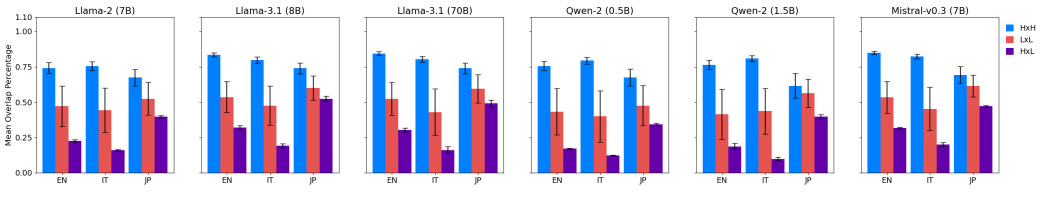
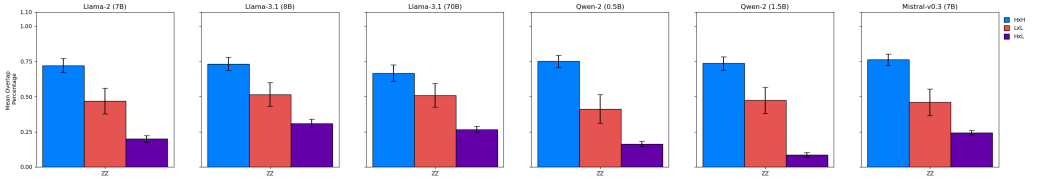


Figure 6: **Experiment 1.** Model-wise accuracy on the grammaticality judgments task given hierarchical and linear Jabberwocky inputs (See § 3.1 and Tables 6 and 3)



(a) English, Italian, and Japanese grammars



(b) Jabberwocky grammars.

Figure 7: **Experiment 2.** Mean pairwise neuron overlaps, by model, for the top 1% of MLP and attention neurons by $\hat{IE}$ between hierarchical and linear inputs. (See § 3.2)

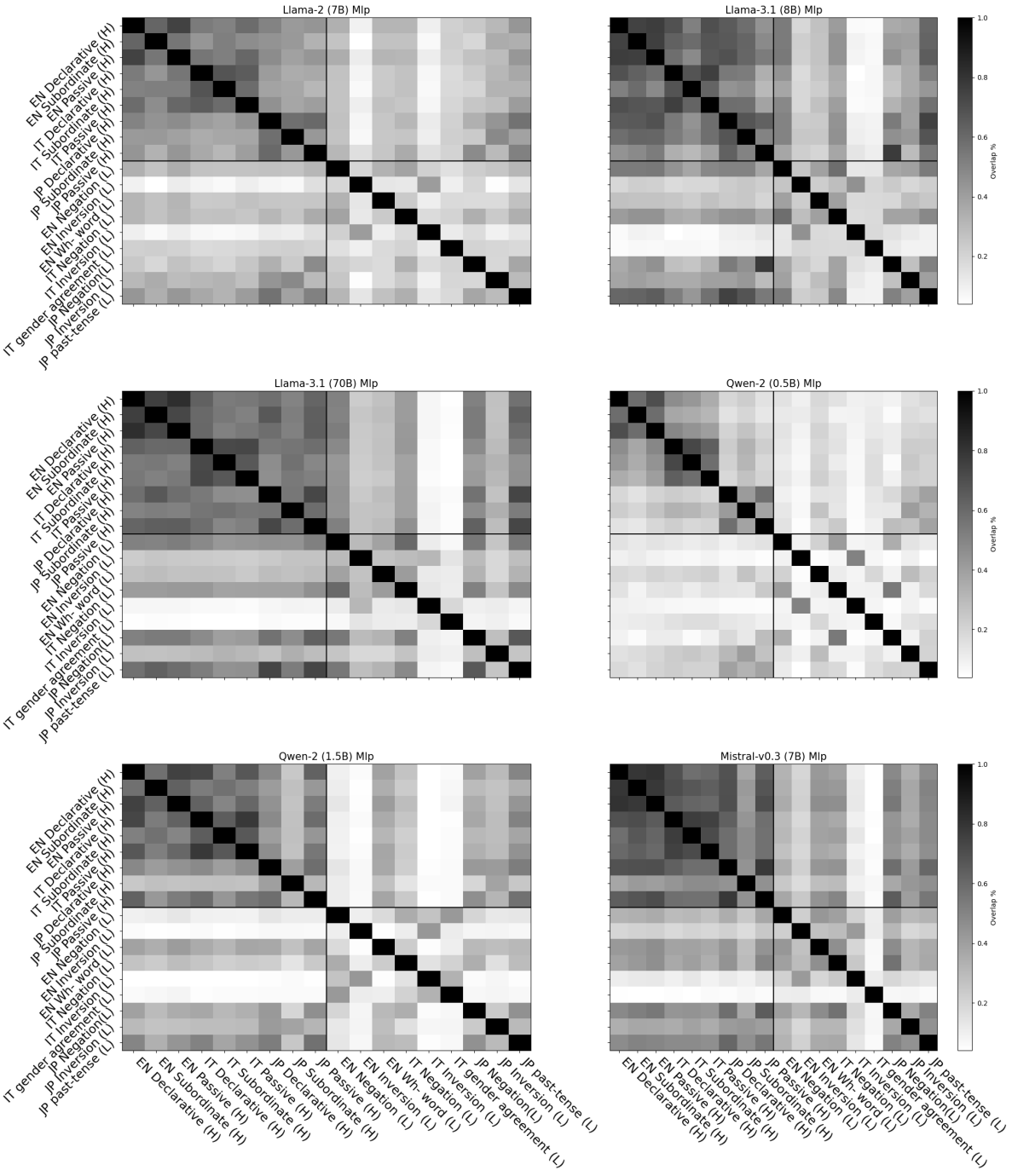


Figure 8: **Experiment 2.** MLP neuron overlaps by model for English, Italian and Japanese grammars.

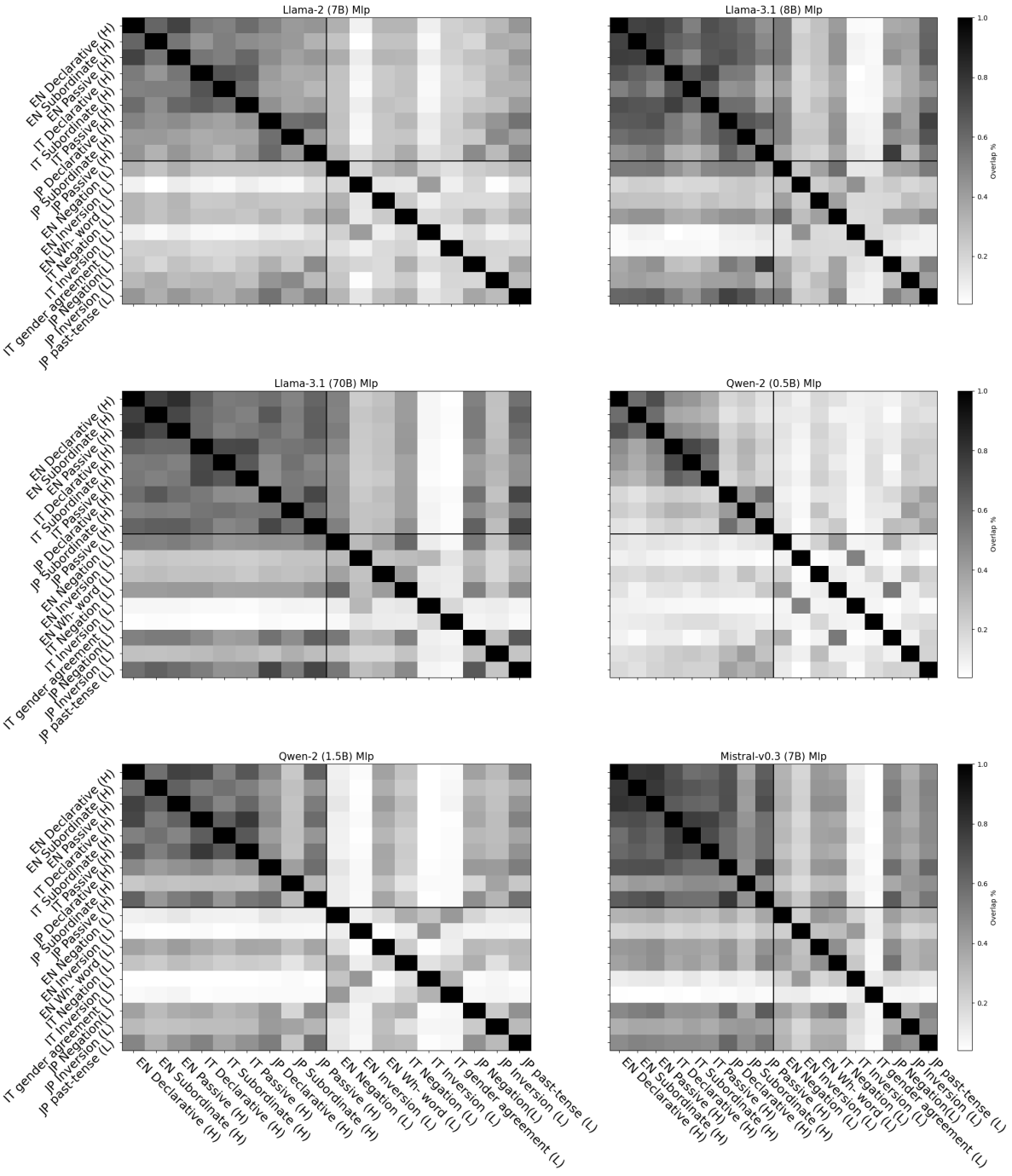


Figure 9: **Experiment 2.** Attention neuron overlaps by model for English, Italian and Japanese grammars.

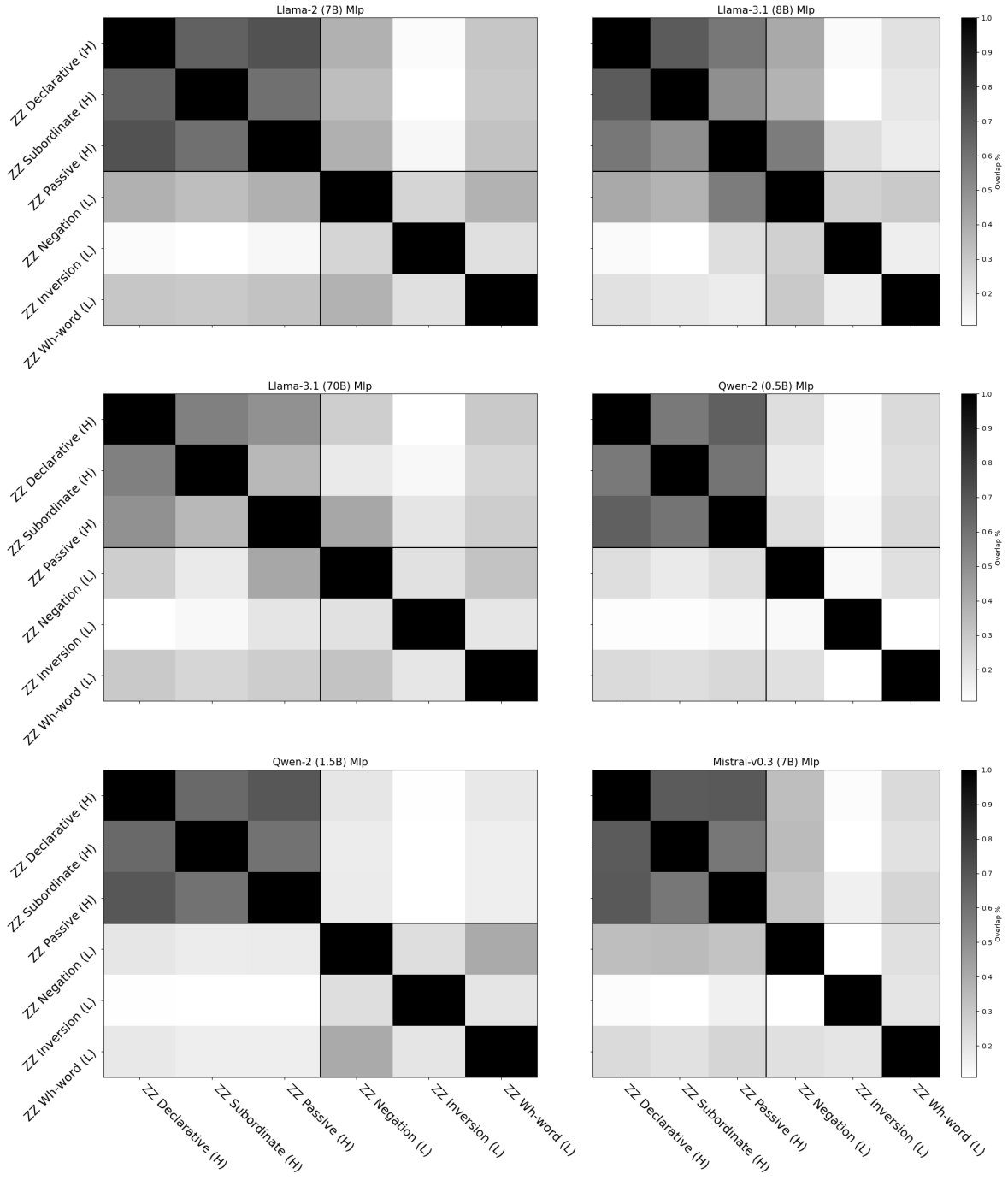


Figure 10: **Experiment 2.** MLP neuron overlaps by model for Jabberwocky grammars.

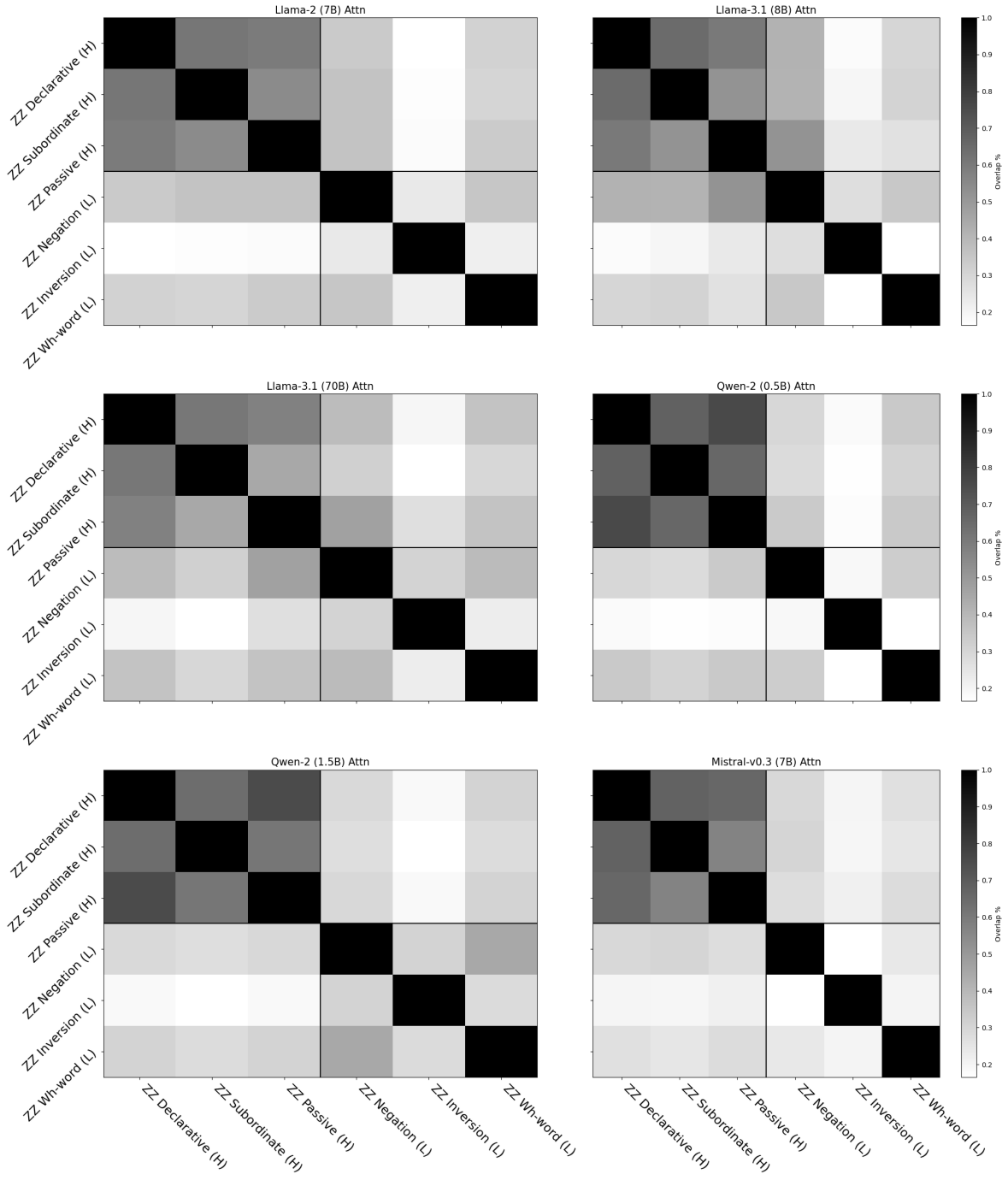


Figure 11: **Experiment 2.** Attention Overlaps by model for Jabberwocky grammars.

Language	Test-Statistic	P-value
EN	5.92	$p < 0.001$
IT	271.5	$p < 0.001$
JP	203	0.2

Table 4: **Experiment 1.** Results from a Mann-Whitney U-Test testing if the model accuracy on the grammaticality judgment task of hierarchical inputs is significantly different from that on linear inputs, when including English, Italian, and Japanese hierarchical and linear inputs ($p < 0.05$).

Grammar	Llama-2-7B	Llama-3.1-8B	Llama-3.1-70B	Qwen-2-0.5B	Qwen-2-1.5B	Mistral-v0.3
EN Declarative (H)	0.80	0.94	0.96	0.68	0.85	0.91
EN Subordinate (H)	0.68	0.92	0.98	0.67	0.76	0.83
EN Passive (H)	0.83	0.93	0.96	0.70	0.78	0.87
EN Negation (L)	0.76	0.83	0.81	0.65	0.39	0.60
EN Inversion (L)	0.65	0.55	0.61	0.69	0.65	0.62
EN Wh-word (L)	0.61	0.61	0.62	0.54	0.55	0.63
IT Declarative (H)	0.74	0.81	0.89	0.63	0.90	0.76
IT Subordinate (H)	0.64	0.71	0.78	0.52	0.87	0.69
IT Passive (H)	0.73	0.84	0.82	0.66	0.87	0.73
IT Negation (L)	0.73	0.87	0.80	0.60	0.60	0.85
IT Inversion (L)	0.61	0.53	0.52	0.59	0.46	0.50
IT Gender Agreement (L)	0.48	0.48	0.50	0.51	0.38	0.44
JP Declarative (H)	0.72	0.95	0.99	0.54	0.67	0.78
JP Subordinate (H)	0.68	0.72	0.80	0.57	0.65	0.58
JP Passive (H)	0.83	0.99	0.98	0.59	0.71	0.90
JP Negation (L)	0.63	0.94	0.99	0.61	0.64	0.75
JP Inversion (L)	0.62	0.65	0.63	0.55	0.61	0.63
JP Past-tense (L)	0.73	0.84	0.99	0.50	0.64	0.58

Table 5: **Experiment 1.** Model accuracies on the grammaticality judgment task for English, Italian, and Japanese hierarchical and linear inputs.

Grammar	Llama-2-7B	Llama-3.1-8B	Llama-3.1-70B	Qwen-2-0.5B	Qwen-2-1.5B	Mistral-v0.3
Declarative (H)	0.68	0.64	0.70	0.57	0.69	0.61
Subordinate (H)	0.64	0.62	0.61	0.53	0.63	0.58
Passive (H)	0.71	0.77	0.89	0.60	0.71	0.62
Negation (L)	0.70	0.73	0.70	0.54	0.45	0.54
Inversion (L)	0.66	0.57	0.59	0.59	0.60	0.52
Wh-word (L)	0.55	0.50	0.55	0.46	0.45	0.48

Table 6: **Experiment 1.** Model accuracies on the grammaticality judgment task for Jabberwocky grammars.

Components	(Test-Statistic, P-value)
H-H vs L-L	(8424, $p < 0.001$)
H-H vs H-L	(11594, $p < 0.001$)
L-L vs H-L	(7792, $p < 0.001$)

Table 7: **Experiment 2.** Results from a Mann-Whitney U-Test investigating whether the overlap percentages for different components across 7 models is significantly different for Jabberwocky grammars. We compare distributions of the mean overlap percentages for the top 1% of MLP and attention components ($p < 0.05$, $N = 108$).

Language	Components	Test Statistic	P-value
English	H-H vs L-L	74794.0	$p < 0.001$
	H-H vs H-L	101160.0	$p < 0.001$
	L-L vs H-L	65649.0	$p < 0.001$
Italian	H-H vs L-L	74794.0	$p < 0.001$
	H-H vs H-L	101160.0	$p < 0.001$
	L-L vs H-L	65649.0	$p < 0.001$
Japanese	H-H vs L-L	74794.0	$p < 0.001$
	H-H vs H-L	101160.0	$p < 0.001$
	L-L vs H-L	65649.0	$p < 0.001$

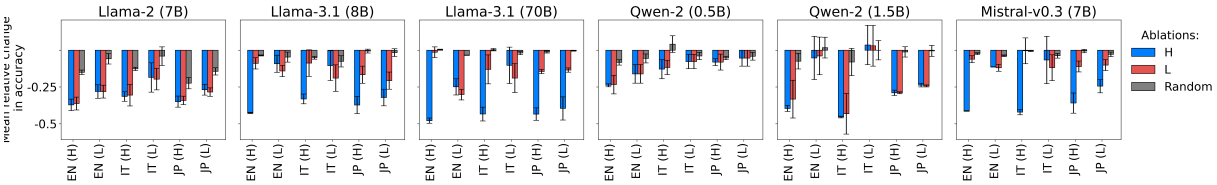
Table 8: **Experiment 2.** Results from a Mann-Whitney U-Test investigating whether the overlap percentages for different components across 7 models are significantly different for English, Italian, and Japanese grammars. We compare distributions of the mean overlap percentages for the top 1% of MLP and attention components ($p < 0.05$, $N = 108$).

Ablation on	**Ablation comparisons (Top 1% of components)**	**Test-statistic**	**P-Value**
EN (H)	H-components vs L-components	31.5	< 0.001
EN (H)	H-components vs R-components	0.0	< 0.001
EN (H)	L-components vs R-components	76.5	0.01
EN (L)	H-components vs L-components	173.5	0.73
EN (L)	H-components vs R-components	52.5	< 0.001
EN (L)	L-components vs R-components	40.0	< 0.001
IT (H)	H-components vs L-components	65.0	< 0.001
IT (H)	H-components vs R-components	15.5	< 0.001
IT (H)	L-components vs R-components	76.5	0.01
IT (L)	H-components vs L-components	193.0	0.33
IT (L)	H-components vs R-components	155.5	0.85
IT (L)	L-components vs R-components	91.5	0.03
JP (H)	H-components vs L-components	77.0	0.01
JP (H)	H-components vs R-components	17.0	< 0.001
JP (H)	L-components vs R-components	41.0	< 0.001
JP (L)	H-components vs L-components	102.0	0.06
JP (L)	H-components vs R-components	25.0	< 0.001
JP (L)	L-components vs R-components	43.5	< 0.001

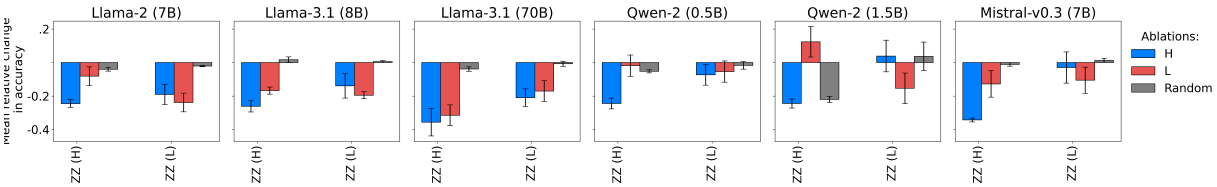
Table 9: **Experiment 3.** Results from a Mann-Whitney U-test investigating whether ablating uniformly sampled model components, as well as components used in the grammaticality judgment tasks of hierarchical and linear grammars containing natural language lexicons significantly differ in how they suppress performance on English, Italian, and Japanese hierarchical and linear grammars. Mean ablations are applied from the structure being judged in the task.

Language	Ablation	Test-statistic	P-Value
ZZ (H)	ZZ(H) vs ZZ(L)	33.0	< 0.001
ZZ (H)	ZZ(H) vs ZZ(R)	14.5	< 0.001
ZZ (H)	ZZ(L) vs ZZ(R)	131.5	0.34
ZZ (L)	ZZ(H) vs ZZ(L)	197.0	0.27
ZZ (L)	ZZ(H) vs ZZ(R)	84.5	0.01
ZZ (L)	ZZ(L) vs ZZ(R)	44.5	< 0.001

Table 10: **Experiment 3.** Results from a Mann-Whitney U-test investigating whether ablating uniformly sampled model components, as well as components used in the grammaticality judgment tasks of jabberwocky hierarchical and linear grammars significantly differ in how they suppress performance on jabberwocky hierarchical and linear grammars. Mean ablations are applied from the structure being judged in the task.



(a) English, Italian, and Japanese grammars



(b) Jabberwocky grammars.

Figure 12: **Experiment 3.** Mean relative change in accuracy by model, when ablating the top 1% of attention and MLP neurons by $\hat{IE}$ between hierarchical and linear inputs. (See § 3.3)

Ablation	Structure Type	Test-statistic	P-Value
EN (H) vs EN (L)	ZZ (H)	69.5	0.004
EN (H) vs EN (Random)	ZZ (H)	11.0	< 0.001
EN (L) vs EN (Random)	ZZ (H)	68.5	0.003
EN (H) vs EN (L)	ZZ (L)	188.5	0.41
EN (H) vs EN (Random)	ZZ (L)	104.0	0.07
EN (L) vs EN (Random)	ZZ (L)	87.5	0.02

Table 11: **Experiment 4.** Results from a Mann-Whitney U-test investigating whether ablating uniformly sampled model components, as well as components used in the grammaticality judgment tasks of hierarchical and linear English grammars significantly differ in how they suppress performance on jabberwocky hierarchical and linear grammars.

Components	(Test-Statistic, P-value)
H(ZZ x EN) vs L(ZZ x EN)	(9678, $p < 0.001$)
H(ZZ x EN) vs H(ZZ) x L(EN)	(11344, $p < 0.001$)
L(ZZ x EN) vs H(ZZ) x L(EN)	(7096, $p < 0.01$)

Table 12: **Experiment 4.** Results from a Mann-Whitney U-Test investigating whether the overlap percentages for different components across 7 models is significantly different for Jabberwocky and English hierarchical and linear grammars. We compare distributions of the mean overlap percentages for the top 1% of MLP and attention components ($p < 0.05$, $N = 108$)

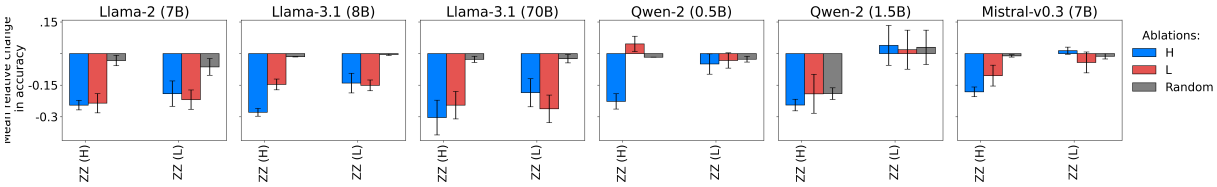


Figure 13: **Experiment 4.** Mean relative change in accuracy by model, when ablating the top 1% of attention and MLP neurons pertaining to English hierarchical and linear grammars. Model is tested on Jabberwocky grammars post ablation, and performance decrease is measured on hierarchical and linear inputs. (See § 3.4)

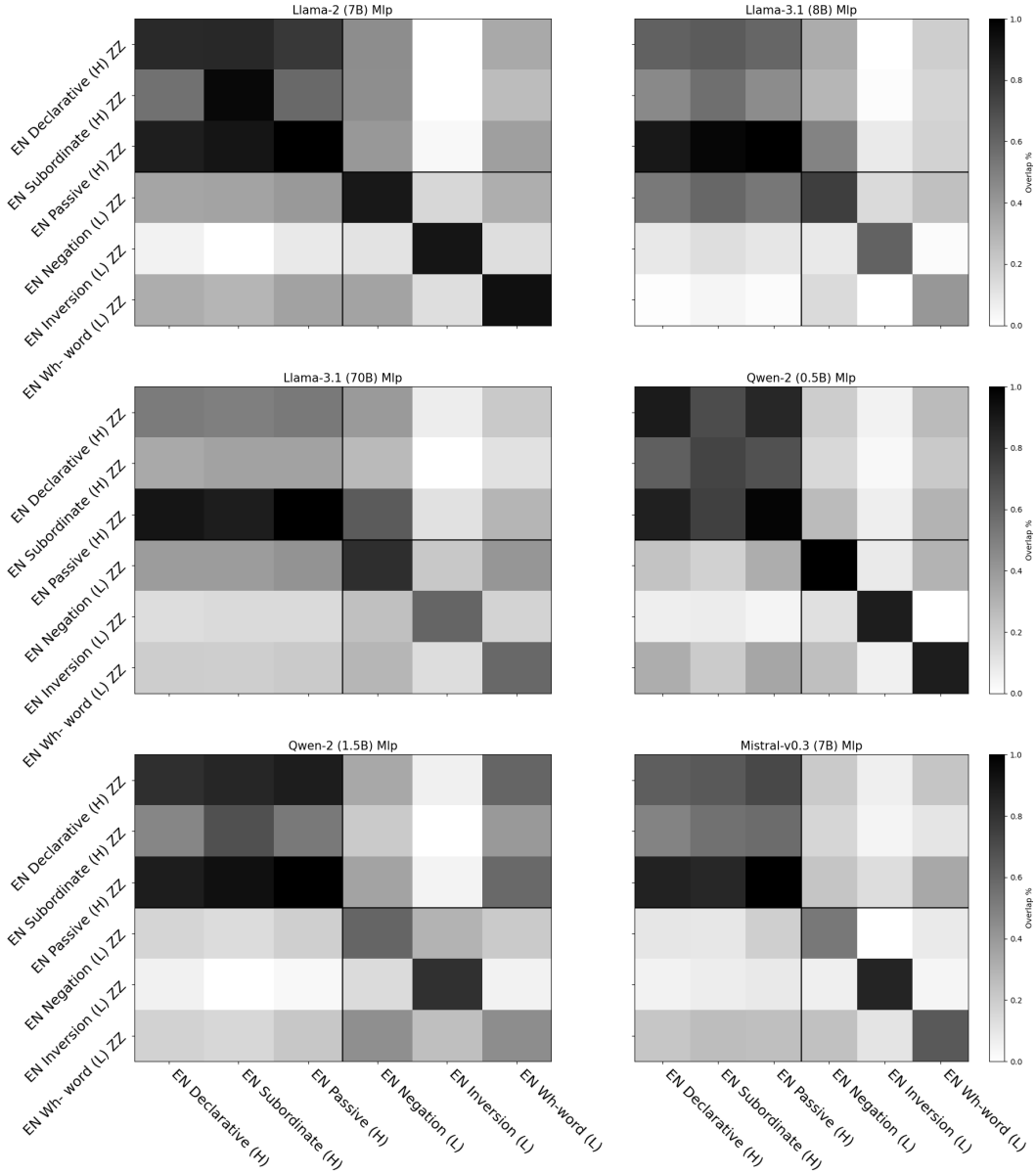


Figure 14: **Experiment 4.** MLP Overlaps by model between English and Jabberwocky grammars

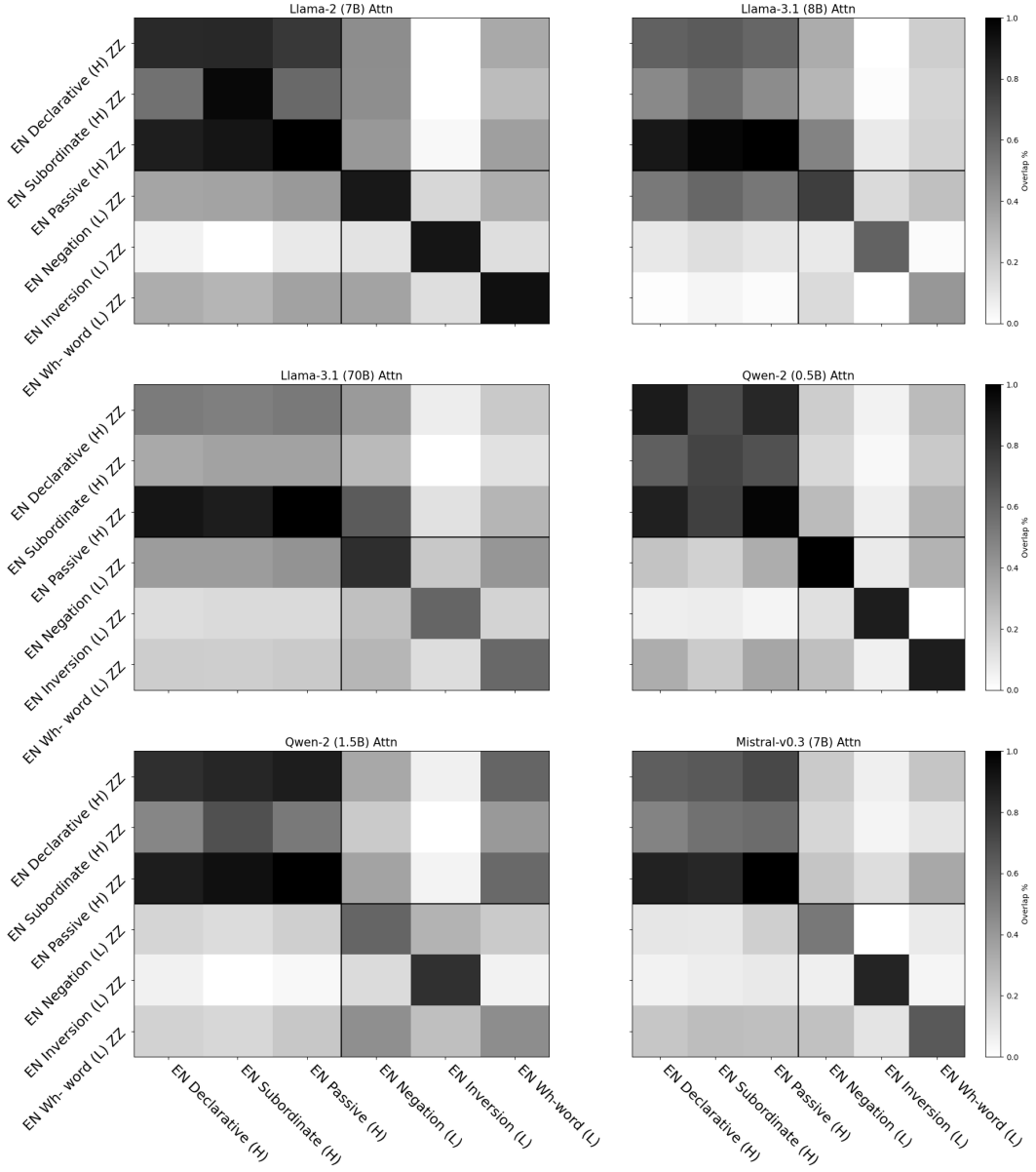


Figure 15: **Experiment 4.** Attention Overlaps by model between English and Jabberwocky grammars