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

Aruna Sankaranarayanan [1] **Dylan Hadfield-Menell** [1] **Aaron Mueller** [2]

Abstract

All natural languages contain hierarchical structure. In humans, this structural restriction is neurologically coded: when presented with linear and hierarchical grammars with identical vocabularies, brain areas responsible for language processing are *only* sensitive to the hierarchical grammar. In this study, we investigate whether such functionally specialized grammar processing regions can emerge in large language models (LLMs) whose processing mechanisms are formed solely from exposure to language corpora. We prompt transformer-based autoregressive LLMs to determine the grammaticality of hierarchical and linear grammars in an in-context-learning setup. First, we discover that models demonstrate higher accuracy, and lower/comparable surprisals, on hierarchical grammars. Next, we use attribution patching to show that model components processing hierarchical and linear grammars are distinct. Lastly, ablating components for hierarchical/linear grammars selectively reduces accuracy for the corresponding grammar. Our findings indicate that large-scale text exposure alone can lead to functional specialization in LLMs.

1. Introduction

In 1861, Broca found evidence that particular cerebral functions were *localized* in specific brain regions, and later discovered evidence for a brain area specialized for language processing (Broca, 1865). 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). The brain's sensitivity toward the structure of natural language (Chomsky, 1957; 1965) is known to be a hallmark of human language processing: natural language is structured hierarchically, and brain regions selective towards hierarchical grammars have been shown to be disjoint from regions selective towards linear structures (Musso et al., 2003), as well as hierarchical but non-linguistic structures such as those found in music or programming languages (Malik-Moraleda et al., 2023; 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). As Malik-Moraleda et al. (2023) notes, "brain areas that process language are *exquisitely* selective for language."

Nonetheless, it is unclear whether language models would demonstrate similar selectivity in the absence of human-like learning biases. Recently, Kallini et al. (2024) find that autoregressive Transformer-based models (Vaswani et al., 2017) more easily acquire grammars that accord with the structures found in human language. While that study investigates language acquisition, we focus on language processing in pre-trained models. We ask two main research questions to understand if language models are sensitive towards hierarchical structures:

1. Do language models demonstrate behavioral distinctions given hierarchical versus linear structures in otherwise identical task settings?
2. Do language models present causally responsible components for judging the grammaticality of hierarchical versus linear inputs across structures and languages?

To answer these questions, we replicate Musso et al.'s experiment on hierarchical and linear selectivity in language processing, to the extent possible,[1] on two large, pre-trained models—Mistral-7B (Jiang et al., 2023) and Llama-70B (Touvron et al., 2023), employing recent causal and mechanistic interpretability techniques to localize model mech-

[1]CSAIL, MIT, Cambridge, MA [2]Northeastern University, Boston, MA. Correspondence to: Aruna Sankaranarayanan <arunas@mit.edu>, Aaron Mueller <aa.mueller@northeastern.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Musso et al. (2003)'s experiment required that subjects be fluent in their native language but not have prior exposure to the foreign languages they were tested on. We cannot guarantee this condition for LLMs, whose training distributions consist of relatively small (but significant) amounts of non-English documents.

anisms. Our results suggest that models are more accurate at processing hierarchical structures, and model regions responsible for processing hierarchical linguistic structure are localizable and distinct. 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, human-like linguistic modularities can emerge.

2. Methods

2.1. Models

We employ the 70B-parameter Llama 2 model(Touvron et al., 2023) (denoted Llama-70B or Llama henceforth) and Mistral-7B v0.1(Jiang et al., 2023) (denoted Mistral-7B or Mistral henceforth) in our experiments. We select these models because they are the best performing open-source models at the time of the experiment (Touvron et al., 2023; Jiang et al., 2023).

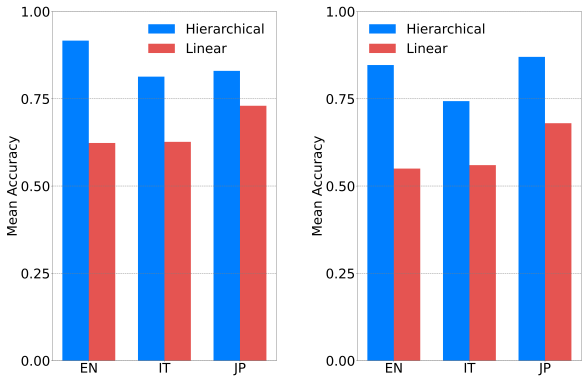
2.2. Data

We define 3 classes of hierarchical and linear grammars in English, Italian, and Japanese. Sentences in each grammar are generated using templates inspired by the constructs defined in (Musso et al., 2003). For each structure, we generate **positive** and **negative** examples. Each grammar, its underlying rule, and examples of corresponding positive and negative examples are available in Table 1. Full descriptions of each grammar template are available in Appendix A. We generate 1106 positive and negative examples for each grammar, totaling to 39816 sentences across all grammars.

The difference between hierarchical and linear grammars lies in whether their latent structure can be explained via positional or hierarchical syntactic rules. Hierarchical grammars contain rules that conform to the structure of natural language, which is hierarchical. Linear grammars contain rules that are defined by word positions or relative word orderings—for example, insert a word at position 4.[2] Such rules are argued to be impossible in human language (Chomsky, 1957; 1965).

3. Experiments

We evaluate Mistral-7B and Llama-70B in an in-context learning setup. Both LLMs are pre-trained on datasets primarily consisting of English sentences.[3]



(a) Mistral-7B (b) Llama-70B

Figure 1: Few-shot accuracies of Mistral-7B (a, left) and Llama-70B (b, right) on hierarchical and linear grammars. We show means across grammars within a language and structural category. For all grammars, both models are significantly ($p < .001$; see Table 3) more accurate on hierarchical structures than linear structures. Individual grammar-level accuracies are provided in Figure 4 in Appendix B.

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

We first evaluate the accuracy of LLMs on grammaticality judgments given examples from each grammar. Our goal from this experiment is to ascertain the behavioural differences in models when processing hierarchical versus linear sentences, as significant behavioural differences may indicate different mechanisms for processing hierarchical and linear grammars. Musso et al. (2003) found that humans were more accurate at classifying examples of hierarchical grammars, even when they had no prior fluency in the test languages; we therefore hypothesize that a similar phenomenon would arise in LLMs if they contain functionally specialized regions for processing hierarchical structure.[4]

We uniformly split the data (§2) in half to obtain our

[2]Note that the surface forms of the sentences are not necessarily linear: in the `EN inversion` sentences, the functional difference is that articles appear after the nouns, which is valid in some languages. Rather, it is the underlying rule that explains each grammar that defines whether it is hierarchical or linear.

[3]89.7% of Llama's pre-training dataset is English; 8.97% is categorized as unknown, with a significant subset of this unknown set being programming code. This suggests that English training data constitutes an even larger percentage of the training data. 0.11% and 0.1% constitute Italian and Japanese sentences, respectively. Mistral-7B is known to be trained on the Open Web, which is also dominated by English text; see here (W3Techs, 2024)).

[4]Note that natural language is largely ambiguous with respect to linear versus hierarchical structure (Chomsky, 1957); human brains have biological preferences for hierarchical structures, but LLMs do not have this preference built into their architecture (Min et al., 2020; McCoy et al., 2018; Chomsky, 1980; Mueller et al., 2022), so it is not clear *a priori* whether they would treat these structures in the same way.

Table 1: **Dataset.** List of grammars, descriptions of the rule defining each grammar, and corresponding positive and negative examples. See Appendix A for full descriptions of each grammar.

	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)

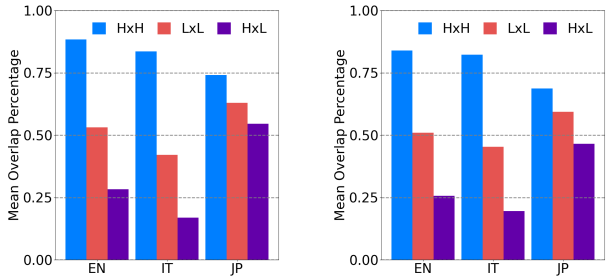
train/test split. Given a structure, we first prompt an LLM with an instruction describing the nature of the in-context-learning task (see Appendix B.1). This is followed by 10 uniformly sampled demonstrations from the training split. The model is then given the metalinguistic judgment task of generating 'Yes' or 'No' when given an example from the test split. Recent research suggests that metalinguistic judgments are inferior to direct probability judgments (Hu and Levy, 2023); to address these concerns, we also measure the difference between positive and negative samples in each model's sentence level surprisal on all hierarchical and linear grammar samples. We report findings from the surprisals task in Figure 5 and in Appendix B.2.

Results. We find (Figure 1, Figure 4, Table 2) that within all languages, Llama-70B and Mistral-7B show higher labeling accuracy ($p < .001$; see Table 3) for sentences from hierarchical as compared to linear grammars. Despite having an order-of-magnitude fewer parameters, Mistral-7B has a higher accuracy than Llama-70B on both hierarchical and linear grammars in English and Italian. We also find that mean surprisal differences are higher for hierarchical grammars in English and Italian, but not Japanese. An independent T-test further shows that differences in surprisal distributions are statistically significant (Table 4) for English and Italian (In Llama, but not Mistral), but not Japanese. These significant behavioral differences suggests a distinction in how these two structures are processed by the model.

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

While §3.1 shows that models are more accurate at judging the grammaticality of hierarchical grammars, it is unclear if the model has specialized and distinct mechanisms for processing hierarchical and linear grammars. To investigate, we locate attention and MLP components that are most sensitive towards processing hierarchical and linear syntax, and test

whether these components have significant overlap. Given our prompt (see Appendix B.1), we quantify the importance of each attention or MLP component, z, in increasing the logit difference m between the correct and incorrect labels for a test sentence t. We do this by estimating the component z's indirect effect (IE; Pearl, 2001; Robins and Greenland, 1992) on y given a test sentence, t, and a minimally different sentence, t' that flips the prediction.[5], using a linear approximation to activation patching (Vig et al., 2020; Finlayson et al., 2021; Geiger et al., 2020; Meng et al., 2022) called **attribution patching** (Kramár et al., 2024; Syed et al., 2023). We then identify the top 1% of model components by IE from the MLP and attention. If there are distinct mechanisms for processing hierarchical and linear grammars, we expect significant overlap between components processing hierarchical-hierarchical and linear-linear grammars, but not hierarchical-linear grammars.



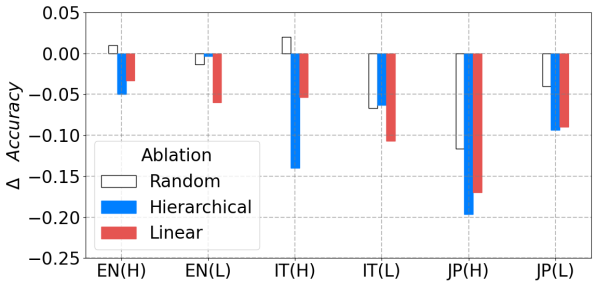
(a) **Mistral-MLP** (b) **Mistral-Attention**

(c) **Llama-MLP** (d) **Llama-Attention**

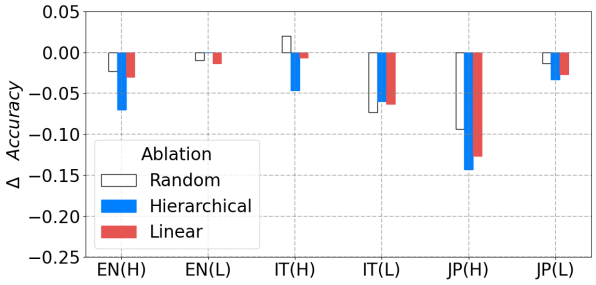
Figure 2: Mean pairwise overlap percentage of the top 1% of MLP and attention components in Mistral-7B and Llama-70B. Overlaps are significantly ($p < .01$; see Table 5) higher for hierarchical-hierarchical pairs than linear-linear pairs, and for hierarchical-hierarchical pairs than hierarchical-linear pairs across languages.

Results. We find (Figure 2) that MLP and attention components in Mistral-7B and Llama-70B show significantly higher ($p < 0.01$) overlap for pairs of hierarchical grammars than pairs of linear grammars (See Table 5) across languages. Further, we find that the mean overlap percentages between hierarchical and linear grammar pairs within a language are significantly lower than those between hierarchical-hierarchical pairs (See Table 5 in App. C). This suggests that large language models contain localized components responsible for processing hierarchical syntax, that are distinct from those responsible for processing linear syntax; this suggests that different parts of the model's computation are dedicated to processing different types of input structures. 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 6,7). This serves as a sanity check that the component overlaps correlate (at least somewhat) based on structural similarities.

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



(a) **Mistral-7B**



(b) **Llama-70B**

Figure 3: Change in accuracies after ablating the top 1% of attention and MLP components of the Mistral-7B and Llama-70B models. For Mistral, ablating hierarchy-sensitive components leads to a significant ($p < .05$) reduction in accuracy for hierarchical grammars, and ablating linearity-sensitive components leads to a (not always significant) reduction in accuracy for linear grammars.

In Experiment 2, we found components responsible for processing hierarchical and linear grammars. Here, we design an ablation experiment to investigate the causal contribution

[5]If t is a positive example, then t' is the corresponding negative example formed by swapping the appropriate words or modifying the sentence, as depicted in Table 1. If t is a negative example, then t' is the corresponding positive example.

of structure-selective model components on grammaticality judgment performance. We first cache the mean activation for each MLP and attention output dimension over the set of training examples[6]. Then, we run three additional iterations of the grammaticality judgment task (§3.1) after: (i) Ablating the union of the top 1% of neurons by $\hat{\text{IE}}$ for all hierarchical grammars, H. (ii) Ablating the sub-sampled set (same size as H) of the top 1% of neurons for all linear grammars, L (sub-sampling procedure described in App.D.1). (iii) Ablating a uniform sub-sample of neurons, where the number of ablated neurons is the same as in (i) and (ii).

Results. For Mistral and Llama (See Figure 3a and 3b), ablating components from H leads to a higher decrease on the model's accuracy on hierarchical structures than linear structures, and vice-versa when ablating components from L. Accuracy differences when ablating uniformly sampled components causes a smaller (if any) decrease in accuracy compared to ablations from H/L (Except in the case of Italian linear structures in Llama). These results hold for all linear/hierarchical structures in English/Italian, and for all hierarchical structures in Japanese. The results hold for 2 out of 3 linear structures in Japanese in Mistral, and none of the 3 structures in Llama. Overall, our results suggest that, for Mistral and Llama, the components discovered in §3.2 selectively reduce model performance depending on structure type for English and Italian. We also conduct Chi-Square tests to investigate if the accuracy differences are statistically significant (See section D.1 and Table 6, and find that the results are mixed, i.e. selective ablations of hierarchical and linear components results in significantly different accuracies (as compared to the no ablations case) 75% of the time, while ablations of randomly sampled components results in significantly different accuracies (when compared to the baseline case) 67% of the time. This adds additional causal evidence to the hypothesis that hierarchical and linear processing is separate in LLMs, and that certain components are functionally specialized toward processing one or the other.

4. Discussion

We find behavioral and causal evidence supporting the hypothesis that hierarchical and linear grammars are processed using largely disjoint mechanisms in LLMs. Thus, as in humans (Baker et al., 2007), general-purpose learners such as neural language models can *acquire* functionally specific regions for the processing of valid linguistic structure. These as well as our behavioral results extend prior evidence that pre-training induces preferential reliance on syntactic features over positional features (Mueller et al., 2022; Murty et al., 2023; Ahuja et al., 2024).

Hierarchical grammars may also be easier to learn than grammars that do not occur in human languages (Kallini et al., 2024; Ahuja et al., 2024). This could explain why language models are so attuned to this structure and learn to explicitly represent it. 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 these syntax-sensitive components and downstream performance.

We use causal localization in our experiments. While not equivalent to explanation, localization can reveal distinctions in where and how certain phenomena are encoded in activation space. Future work could employ other techniques from the training dynamics and interpretability literature to better understand how and when these components arise during pre-training, as well as the functional sub-roles of these distinct component sets.

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. Secondly, it is unclear if the results we observe are due to the model's selective processing of hierarchical versus linear structures or in-distribution vs out-of-distribution structures since the model has only been exposed to hierarchical structures in the training data. Thirdly, our experiments have so far reflected results from a single run of experiments on the model. We expect to repeat these experiments using different random seeds, to improve the robustness of the results. Finally, 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; therefore, we cannot confidently conclude that the significant accuracy differences we observe will generalize to other task settings or prompt formats given the same grammars.

[6]We note that unlike the setup in (Meng et al., 2022), our prompts and test inputs have varying lengths. A given token position doesn't have inherent meaning in our setup. Therefore we aggregate activations across positions

Acknowledgements

AS would like to acknowledge travel grants from MAIA, BTL, and IBM Research, for supporting her travel to ICML 2024. The authors would like to thank Dr. Forrest Davis, Jaden Fiotto-Kaufmann, members of the Baulab, and members of the AI Alignment group for their valuable comments, and implementation help, at various stages of this project. AS would like to thank Carmelo Presicce and Dr. Takako Aikawa for proof-reading the Italian and Japanese sentences generated in our dataset. Finally, AS would like to thank Prof. Robert Berwick, for introducing her to the fascinating world of impossible grammars and the original Musso paper, and Prof. David Bau for introducing her to AM – this paper would not have been possible without them.

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically. *Computing Research Repository*, arXiv:2404.16367, 2024.

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

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

Chris I. Baker, Jia Liu, Lawrence L. Wald, Kenneth K. Kwong, Thomas Benner, and Nancy Kanwisher. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 104(21):9087–9092, 2007. doi: 10.1073/pnas.0703300104. URL https://www.pnas.org/doi/abs/10.1073/pnas.0703300104.

Paul Broca. 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, 1861.

Paul Broca. Sur le siége de la faculté langage articulé. *Bulletins et mémoires de la Société d'anthropologie de Paris*, 6:337–393, 1865.

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

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

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

Noam Chomsky. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15, 1980.

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. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Evelina Fedorenko and Rosemary Varley. 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, 2016.

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

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. *arXiv preprint arXiv:2106.06087*, 2021.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. *arXiv preprint arXiv:2004.14623*, 2020.

Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18, 2018.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. URL https://aclanthology.org/N01-1021.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.306. URL https://aclanthology.org/2023.emnlp-main.306.

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

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

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*, 2024.

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

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

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

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

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

R Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*, 2018.

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

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*, 2020.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, 2012.

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

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. Coloring the blank slate: Pretraining imparts a hierarchical inductive bias to sequence-to-sequence models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.106. URL https://aclanthology.org/2022.findings-acl.106.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. Grokking of hierarchical structure in vanilla transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.38. URL https://aclanthology.org/2023.acl-short.38.

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

Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2023.

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

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992. ISSN 10443983. URL http://www.jstor.org/stable/3702894.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.230. URL https://aclanthology.org/2021.emnlp-main.230.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

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

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

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

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

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

W3Techs. Usage statistics and market share of content languages for websites, may 2024, 2024. URL https://w3techs.com/technologies/overview/content_language. Accessed: 2024-05-18.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470, 2023.

A. Grammar rule descriptions

We define 3 classes of hierarchical sentences in English, Italian, and Japanese.

- `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 subject of a transitive verb in a prepositional phrase. We use the verb and object without the subject, since the use of the subject is not a strict requirement in Italian.[7]

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

- `Negation`: We break the hierarchical order by inserting a negation word "doesn't" 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.

[7]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 third construction varies between languages.

- *Italian*: `Last-noun agreement`: We change the subject term's gender to always match that of the final noun in the sentence.
- *English*: `Wh- word`: 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.
- *Japanese*: `Past Tense`: 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.

B. Experiment 1: Few-shot learning accuracy

Experiment 1 assesses the model's accuracy on grammaticality judgments of hierarchical and linear structures. Here we share (i) examples of prompts used in the experiment, (ii) statistical comparisons of the accuracy distributions in (Table 3) and (iii) grammar-wise accuracy in Figure 4

B.1. Example Prompts

Here are some example prompts from hierarchical structures. The prompt skeleton is in English, irrespective of the language of the examples. Additionally, note that we intentionally strip the whitespace after the final `A:`, to aid consistent label generation. This was particularly important when deploying experiments on Llama-70B which was prone to generating streams of whitespace characters if we did not strip the final whitespace in the input prompt.

B.1.1. 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:"
```

B.1.2. 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
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:"
```

B.1.3. 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 1b: How do models' surprisals compare when processing hierarchical vs linear structures?

Both Mistral-7B and LLama-70B demonstrate higher accuracy on hierarchical grammars. However, it is unclear whether this is due to hierarchically structured inputs being more predictable for the model. (Musso et al., 2003) had participants press a button to mark a sentence as being grammatical or not, based on the initial grammar template that they were exposed to. They measured reaction times (RT)—i.e., the time taken for a participant to process a sentence and make a decision. Their experiment found that reaction times were higher when participants were judging sentences from hierarchical grammatical constructions as compared to linear constructions. Following Hale (2001), Levy (2008),Monsalve et al. (2012), Goodkind and Bicknell (2018), Wilcox et al. (2023), Oh and Schuler (2023), we use token surprisals from a language model as a proxy for human reading times, which we then use as a proxy for the LLM's reaction time in this classification task.

We compare the sentence-level surprisals across hierarchical and linear structures. The token-level surprisal $S(t_n)$ of the n^{th} token t_n is the negative log-probability of t_n given context $t_1, ..., t_{n-1}$:

$$S(t_n) = 1 - \log p(t_n | t_1, ..., t_{n-1}) \quad (1)$$

The sentence-level surprisal $S(T)$ is the sum of the token-level surprisals for all tokens in the sentence.

$$S(T) = \sum_{i=1}^{n} S(t_i) \quad (2)$$

For each example, we compute the difference in surprisals $S(T) - S(T')$, where T is the positive example (corresponding to a sentence following the grammatical rule) and T' is the negative example; thus, if the model behaves as expected, surprisals should be negative, as it should assign higher probability to the positive example. Then, we compute the mean of this difference across examples within a language and structural category (hierarchical or linear).

We hypothesize that the surprisal differences will be lower for hierarchical grammars than linear grammars if the model is specialized toward processing hierarchical grammars. If the model can effectively process linear inputs with equal ability, then surprisal differences should be comparable.

Results. We observe (Figure 5) that surprisal differences between hierarchical and linear grammars are significant for English for Llama ($p < .05$; see Table 4), but not for other languages or for any pair in Mistral. These findings are different from the findings with respect to reading speed as shown in (Musso et al., 2003), where humans have better reading speeds on hierarchical grammars as compared to linear grammars over time; they also differ from past findings with respect to surprisals during languqge acquisition by the model (Kallini et al., 2024), where it is found that the surprisal is significantly lower for hierarchical than linear grammars. Thus, though large pre-trained models are predominantly conditioned on hierarchical inputs, they do not always predict linear structures as being significantly more unlikely in this task setting. This suggests that models might have coherent ways of processing linear inputs, or at least do not have significantly lower confidence in processing them.

C. Experiment 2: MLP and Attention Component Overlaps

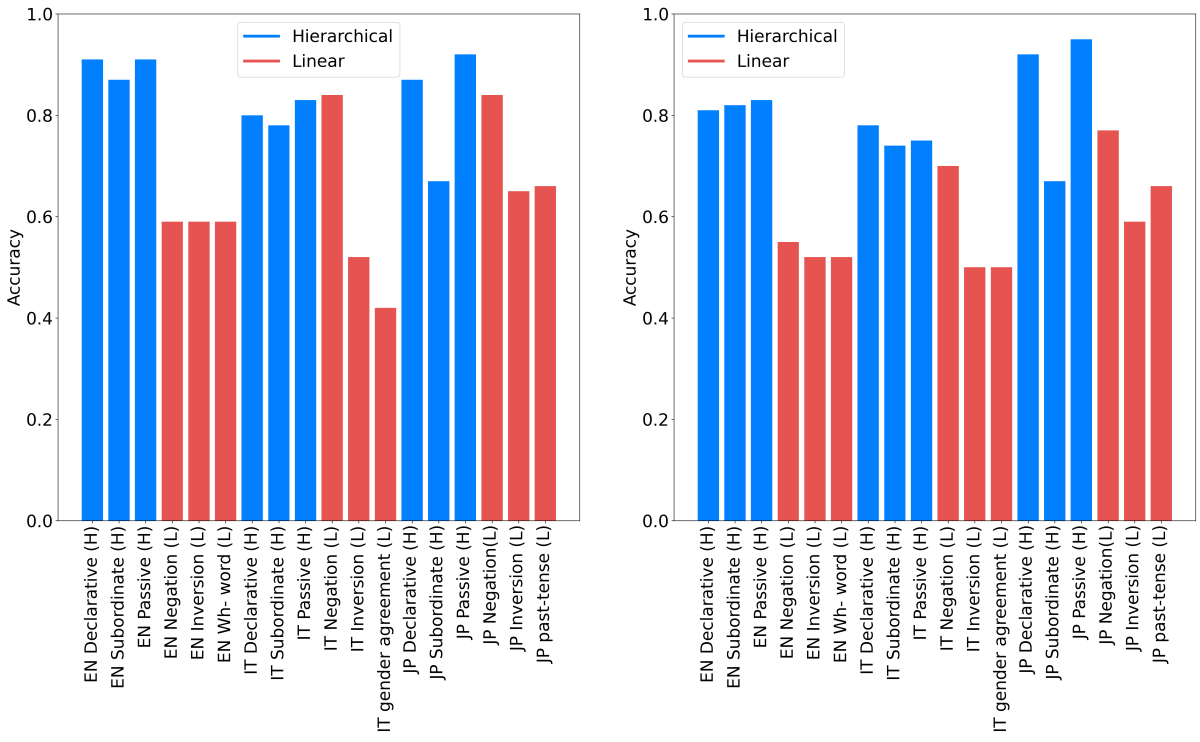
Experiment 2 locates MLP and Attention components involved in processing hierarchical and linear structures and investigates if these components are disjoint. §C.1 describes the attribution patching algorithm. The grammar level pairwise overlaps for all grammars across EN, IT, and JP is given in Figures 7 and 6. Table 5 shows the statistical tests for comparing overlaps between pairs of grammar structures.

Table 2: **Experiment 1:** Few-shot classification accuracy of Llama and Mistral over our dataset of hierarchical and linear grammars.

	Mistral		Llama	
Structure	Hierarchical	Linear	Hierarchical	Linear
EN	0.92	0.62	0.85	0.55
IT	0.81	0.63	0.74	0.56
JP	0.83	0.73	0.87	0.68

Table 3: **Experiment 1:** Statistical comparisons of the accuracy distributions of Llama and Mistral models on a grammaticality judgment task. We compare the distributions of the model's accuracy on predicting the grammaticality of hierarchical and linear grammars, using both an independent samples t-test (where we consider the accuracies as being either a 1 or 0) and a Chi-Square test (where we consider the accuracies as being either True or False) along with a Bonferroni correction. Note that N=3316, and the adjusted p-value threshold = 0.05/3 = 0.016.

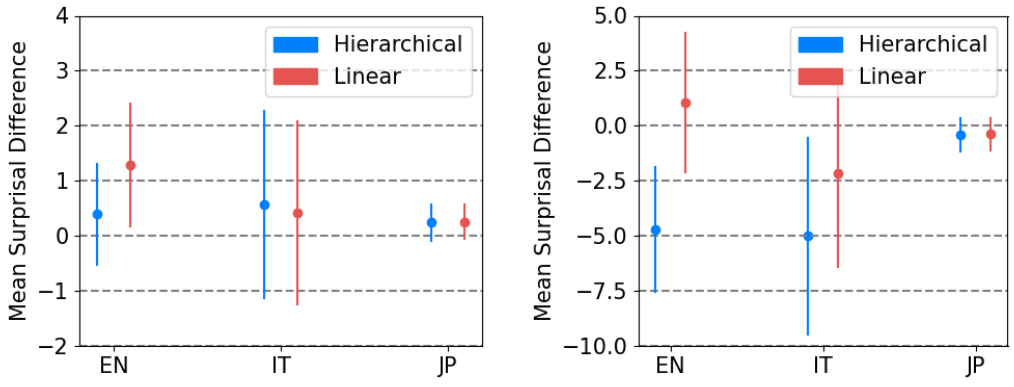
Language	**Mistral**		**Llama**	
	Chi-Square Test	**t-Test**	**Chi-Square Test**	**t-Test**
EN	$(408.113, p < 0.001)$	$(21.616, p < 0.001)$	$(338.285, p < 0.001)$	$(19.447, p < 0.001)$
IT	$(147.666, p < 0.001)$	$(12.469, p < 0.001)$	$(120.378, p < 0.001)$	$(11.211, p < 0.001)$
JP	$(48.066, p < 0.001)$	$(7.024, p < 0.001)$	$(171.318, p < 0.001)$	$(13.482, p < 0.001)$



(a) Mistral-7B

(b) Llama-70B

Figure 4: Structure-wise few-shot accuracies of Mistral-7B (a, left) and Llama-70B (b, right) on a labeling task involving hierarchical and linear grammar inputs. Note that task-accuracy on hierarchical grammar inputs is consistently higher than that on linear grammar inputs.



(a) Mistral-7B

(b) Llama-70B

Figure 5: Surprisal differences between positive and negative examples of hierarchical and linear sentences. The bars represent the 95% confidence interval, and the dot represents the *mean* surprisal across examples. The difference between the distributions is significant for hierarchical and linear structures in English for Llama, but not significant for Italian or Japanese, and for any structures processed by Mistral.

Table 4: Statistical comparisons of the surprisal distributions of Llama and Mistral models on sentences from the grammaticality judgment task using an independent t-test. The values are represented as (t-statistic, p-value). p-value=0.001.

Language	Mistral (t-Test)	Llama (t-Test)
EN	(-1.2, $p = 0.22$)	(-2.66, $p < 0.05$)
IT	(0.12, $p = 0.90$)	(-0.91, $p = 0.36$)
JP	(-0.05, $p = 0.96$)	(-0.07, $p = 0.94$)

C.1. Activation and Attribution Patching

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

$$\text{IE}(l; z; t, t') = l(t|\text{do}(z_t = z_{t'})) - l(t) \quad (3)$$

Activation patching is computationally expensive, requiring 2 forward passes for all model components; i.e., we have $O(n)$ forward passes, where n is the number of components we investigate. Therefore, we opt to use attribution patching (Kramár et al., 2024; Syed et al., 2023), a first-order Taylor approximation of the IE that we would have obtained via activation patching:

$$\hat{\text{IE}}(l; z; t, t') = \nabla_z l|_{z=z_t} (z_{t'} - z_t) \quad (4)$$

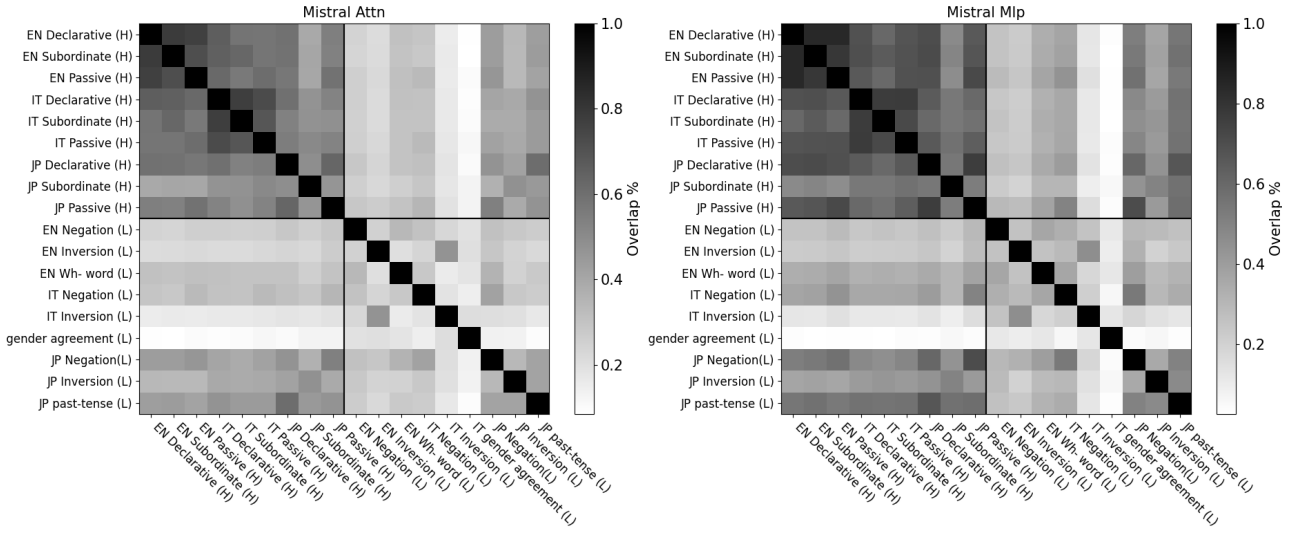
We can measure this quantity using only 2 forward passes and 1 backward pass for all z; i.e., we have $O(1)$ passes. While this approximation is not perfect, it correlates strongly with the actual IE in typical cases (Kramár et al., 2024; Marks et al., 2024). The top 1% of attention and MLP components in the model are selected by $\hat{\text{IE}}$ values. We compute the pairwise overlap percentage of this top 1% neuron subset for each pair of grammars.

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

D.1. Sub-sampling top 1% of MLP and attention neurons for ablation

Ablation neurons in 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 find 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.

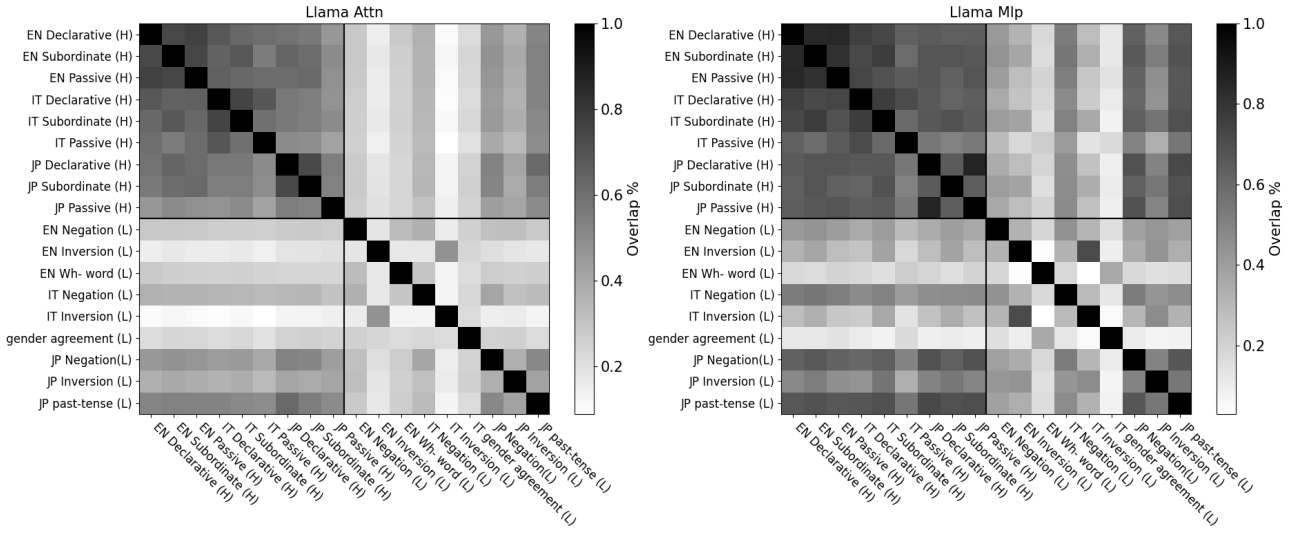
We conduct statistical tests to determine if the accuracy distributions of hierarchical and linear sentences, that were labeled differently with and without ablation, are significantly different. The results are shown in 6. Note that non-random ablations lead to significantly different labeling accuracies for hierarchical and linear grammatical struc-



(a) Mistral Attention

(b) Mistral MLP

Figure 6: **Experiment 2:** Confusion matrix showing the percentage of overlapping neurons in the Mistral-7B model for EN, IT, and JP structures. Hierarchical structures show consistent overlaps as compared to linear structures, particularly for English and Italian.



(a) Llama Attention

(b) Llama MLP

Figure 7: **Experiment 2:** Confusion matrix showing the percentage of overlapping neurons in the Llama-70B model for EN, IT, and JP structures. Hierarchical structures show consistent overlaps as compared to linear structures, particularly for English and Italian.

tures 75% of the time. However, random ablations also cause significantly different behaviours 67% of the time. Therefore, while the ablations result in different labeling accuracies for hierarchical and linear grammatical structures, this difference is not always significant.

Table 5: **Experiment 2:** Statistical comparisons of the neuron overlap percentages between pairs of hierarchical, linear and hierarchical-linear structures, using a Mann-Whitney U test along with a Bonferroni correction. The highlighted row does not show a statistically significant difference. Note that N=9, and the adjusted p-value threshold = 0.05/3 = 0.016

	Llama		Mistral	
	MLP (statistic, p value)	Attention (statistic, p value)	MLP (statistic, p value)	Attention (statistic, p value)
H-H vs L-L	514.5, $p = 0.008$	526.5, $p = 0.004$	526.5, $p = 0.004$	526.5, $p = 0.004$
H-H vs H-L	695.0, $p < 0.001$	717.0, $p < 0.001$	709.0, $p < 0.001$	715.0, $p < 0.001$
L-L vs H-L	419.0, $p = 0.35$	459.0, $p = 0.10$	465.0, $p = 0.08$	459.0, $p = 0.10$

Table 6: **Experiment 3:** Chi-square statistics and p-values for Mistral and Llama models across different languages and ablation types. Note that accuracy differences are significant even when

	Mistral			**Llama**		
Language	**Real**	**Unreal**	**Random**	**Real**	**Unreal**	**Random**
EN	(22.81, $p < 0.001$)	(11.62, $p < 0.001$)	(2.93, 0.09)	(31.32, $p < 0.001$)	(3.17, 0.08)	(1.12, 0.29)
ITA	(12.73, $p < 0.001$)	(1.30, 0.25)	(27.32, $p < 0.001$)	(0.35, 0.55)	(13.82, $p < 0.001$)	(33.86, $p < 0.001$)
JAP	(42.04, $p < 0.001$)	(20.33, $p < 0.001$)	(30.86, $p < 0.001$)	(58.17, $p < 0.001$)	(48.38, $p < 0.001$)	(44.36, $p < 0.001$)