Disaggregation Reveals Hidden Training Dynamics: The Case of Agreement Attraction

James A. Michaelov
MIT
jamic@mit.edu

Catherine Arnett
EleutherAI
catherine@eleuther.ai

Abstract

Language models generally produce grammatical text, but they are more likely to make errors in certain contexts. Drawing on paradigms from psycholinguistics, we carry out a fine-grained analysis of those errors in different syntactic contexts. We demonstrate that by disaggregating over the conditions of carefully constructed datasets and comparing model performance on each over the course of training, it is possible to better understand the intermediate stages of grammatical learning in language models. Specifically, we identify distinct phases of training where language model behavior aligns with specific heuristics such as word frequency and local context rather than generalized grammatical rules. We argue that taking this approach to analyzing language model behavior more generally can serve as a powerful tool for understanding the intermediate learning phases, overall training dynamics, and the specific generalizations learned by language models.

1 Introduction

Though only a recent development (see, e.g., Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Warstadt et al., 2019, 2020; Gauthier et al., 2020; Hu et al., 2020), it is almost taken for granted today that language models tend to generate grammatical strings of text. Indeed, contemporary large language models have been argued to show linguistic competence (Mahowald et al., 2024). But even large models such as Chinchilla have been shown to often fail at more difficult grammatical tasks (Lampinen, 2024), suggesting that rather than learning fully general rules, models may be learning more specific rules or increasingly complex heuristics. In this study, we investigate what generalizations language models *do* learn by turning to two highly influential approaches from the study of human language—analyzing errors (e.g., Fromkin, 1971; Garrett, 1975; Dell, 1986; Bock and Miller, 1991) and studying changes over the course of acquisition (e.g., Kenney and Wolfe, 1972; Rumelhart et al., 1986; Marcus et al., 1992).

In this paper, we focus on subject-verb agreement, which is the fact that in a sentence such as *the cat leaps*, the word *leaps* correctly agrees with the subject *cat*; while *the cat leap* is grammatically incorrect. In the simplest cases, subject-verb agreement appears relatively easy to learn for language models. It is learned early in training (Evanson et al., 2023) and at human-scale levels of training data (Warstadt et al., 2023; Hu et al., 2024; Wilcox et al., 2025), and can be learned even for low-resource languages (Jumelet et al., 2025) and by traditional LSTM-RNNs (Linzen et al., 2016). However, language models appear to struggle at subject-verb agreement in more complex linguistic structures that are also difficult for humans. For example, while language models are good at predicting the correct form of the verb (V) in sentences like (1) based on the subject noun (S), their overall performance drops in cases where there is an intervening attractor noun (A) as in (2) (Marvin and Linzen, 2018; Gulordava et al., 2018; Warstadt et al., 2020; Arehalli and Linzen, 2020; Ryu and Lewis, 2021; Lakretz et al., 2021; Lampinen, 2024), and they are also slower to reach their peak performance on such items (Evanson et al., 2023). The crucial point to note, however, is that language

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

model performance is not uniformly worse on such stimuli—like humans (see, e.g. Bock and Miller, 1991; Bock and Cutting, 1992; Franck et al., 2002), they have been observed to show a higher error rate in cases where there is a mismatching attractor, i.e., in cases such as *the athletes near the bike* ***know/knows*** (Arehalli and Linzen, 2020; Ryu and Lewis, 2021; though see Lampinen, 2024). This may suggest that the models are not making their predictions on the basis of a general subject-verb agreement rule, but rather based on more specific patterns or or surface-level heuristics (for a more general discussion on language model behavior in this vein, see, e.g., McCoy et al., 2024).

$$\text{The } \left\{ \begin{array}{c} \text{athlete} \\ \text{athletes} \end{array} \right\}_{\text{S}} \left\{ \begin{array}{c} \text{knows} \\ \text{know} \end{array} \right\}_{\text{V}} \ldots \tag{1}$$

$$\text{The } \left\{ \begin{array}{c} \text{athlete} \\ \text{athletes} \end{array} \right\}_{\text{S}} \text{near the } \left\{ \begin{array}{c} \text{bike} \\ \text{bikes} \end{array} \right\}_{\text{A}} \left\{ \begin{array}{c} \text{knows} \\ \text{know} \end{array} \right\}_{\text{V}} \ldots \tag{2}$$

We carry out an exploratory analysis looking at how the language model performance on different data subsets—corresponding to different experimental manipulations of sentences of the form shown in (1) and (2)—vary over the course of training, and how these patterns compare to those observed in when looking at the aggregate performance across all sentence types.

2 Method

Datasets We use the `simple_english` and `nounpp_english` (i.e., subject-verb agreement with prepositional phrase attractor) subsets of the Subject-Verb Agreement task (Linzen et al., 2016; Marvin and Linzen, 2018; Gulordava et al., 2018; Goldberg, 2019; Wolf, 2019; Lakretz et al., 2019, 2021) in BIG-bench (Srivastava et al., 2023). We also add to this the corresponding (i.e., subject-verb agreement with prepositional phrase attractor) subset of the stimuli from Bock and Cutting (1992), as preprocessed by Arehalli and Linzen (2020); which differ from the BIG-bench stimuli in that they all only include the verb *to be* (i.e., *is*/*are*), and that in some cases, the subject is more than one word long (e.g., *the teaching assistant*). We additionally create simple agreement sentences by removing the intervening prepositional phrases from these stimuli.

Models We use the PolyPythia suite of language models (van der Wal et al., 2024). These are a set ten random seeds of each Pythia model (Biderman et al., 2023) from 14M to 410M parameters, released with multiple checkpoints over the course of training. Using these models allows us to look at the training dynamics of the models while accounting for differences arising from random initialization and data shuffling. Each training step represents the same number of tokens seen (with the same tokenizer), therefore, we can compare across model sizes and random seeds.

Procedure We calculate the log-probability of the each verb following its context. We consider the model to be correct if it assigns a higher log-probability to the correct form of the verb relative to the incorrect form. For some verbs, the singular and plural forms are each represented by one token; while for others, the singular form (e.g., *admires*) is represented by two tokens and the plural form (e.g., *admire*) by one. We refer to these as single-token and multi-token verbs, respectively, and analyze them separately. We define the log-probability of multi-token words as the sum of their token log-probabilities (as a further analysis, we also consider normalizing by number of tokens; see Appendix D). We release all code in the following repository: `https://github.com/jmichaelov/sv-disaggregation-cognitive-interpretability`.

3 Results

We present the results of our analysis in Figure 1. First, we consider performance as measured by the average accuracy across all conditions. In general, we see little improvement (and in some cases, a drop in accuracy) until steps 256–512, at which point we see an increase in performance, which is smaller and more gradual for the smaller models (and begins later for the Pythia 14M) and faster and more complete for the larger models.

Condition-level accuracy reveals a rather different set of patterns. For the *be* verb (i.e., *is* vs. *are*), we initially see a high accuracy for the singular conditions and a low accuracy for the plural conditions;

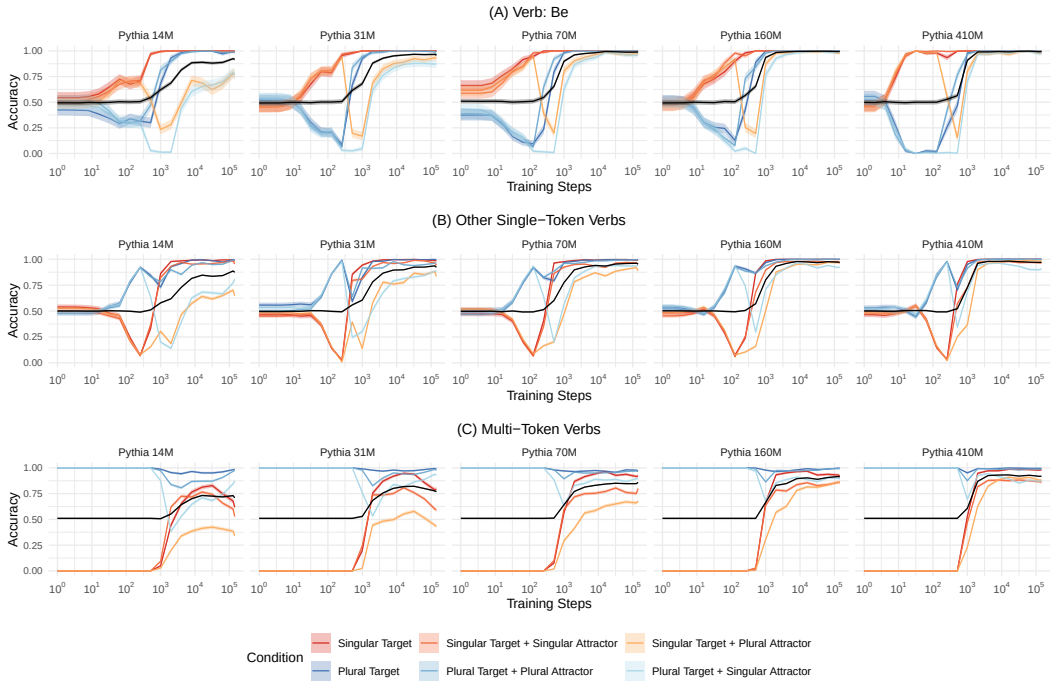

Figure 1: PolyPythia model accuracy on subject-verb agreement stimuli with (A) the verb *be*, (B) all other single-token and (C) multi-token verbs. The black line represents the mean across all conditions (i.e., the aggregate score). Shading reflects 95% confidence intervals.

i.e., the model is virtually always predicting the word *is* to be more likely than *are*. Then, around steps 128-512, we see a sharp increase in accuracy for the plural and plural-with-plural-attractor conditions, but not the plural-with-singular-attractor condition; and we see a corresponding sharp decrease in singular-with-plural-attractor condition—that is, we see the agreement attractor effect. After this, we see an overall increase in performance for all conditions. For the other verbs, we see the reverse pattern—a preference for the plural (e.g., *admire*) over the singular (e.g., *admires*), followed by a drop in performance of the plural-with-singular-attractor condition and an increase in the singular and singular-with-singular-attractor conditions, and finally, an overall increase. Disaggregating the results for each verb (see Appendix B), we see that while most of these patterns are present for each verb and each model, there is some variation—for example, the preference for *stimulate* over *stimulates* is much smaller than for the other verbs. Additionally, in some cases, there appears to be a brief reversal of the singular vs. plural preference at step 512 (for example, with *observe*), though this quickly reverses. There is also some variation by random seed (Appendix C). In general, the patterns are less stable overall in the smaller models.

For multi-token verbs, we see overall the same pattern: an initial (and more immediate) preference for (one-token) plural over (two-token) singular verbs, followed by an increase in accuracy for verbs in the singular and singular-with-singular-attractor conditions and a corresponding decrease in accuracy in the plural-with-singular-attractor condition, followed by overall improvement on all conditions. However, these patterns generally occur later in training, and the performance decrease is smaller.

4 Discussion

In this study, we show that aggregated metrics of performance may hide interpretable patterns in the trajectory of language models' grammatical knowledge over the course of training. We see evidence of clear systematic patterns at both the condition level and based on verb tokenization. In contrast to the overall slow and gradual increase in performance, disaggregation reveals rapid (and often non-monotonic) changes in the behavior that underlie this and that begin far earlier. Initially, models

learn to assign a higher probability to the more frequent form of the verb. In the case of *be*, *is* is more frequent than *are*; and in all other cases, the plural form of the verb (as the bare form of the verb) is more frequent in the training corpus (see Appendix A). Next, the models appear to become sensitive to the preceding word: we see a sharp improvement in performance on the simple condition of the less frequent verb form, as well as the condition with the matching attractor. There is also a decrease in performance on the mismatched attractor condition for the more frequent verb form; i.e., we see a strong effect of the attractors on performance. Finally, performance continues to improve until the end of training. A possible explanation for the first two phases is the finding that over the course of training, transformers overfit their predictions to token unigram probability (i.e., frequency), then bigram probability, then trigram probability, and so on (as described in, e.g., Chang et al., 2024). This may also explain the later-occurring preference for singular verbs following singular nouns (targets or attractors) in multi-token verbs relative to single-token verbs. For single-token verbs, a sensitivity to the number of the preceding noun only requires taking into account the previous token (i.e., bigram-like behavior), but for the second token of a multi-token verb (which only occurs for singular forms of the verb), it requires taking into account the previous two (trigram-like behavior). Whether the models are displaying strictly n-gram-like behavior or a more general ability to make predictions based on an increasingly long context is a question for future work.

In either case, these findings highlight the importance of considering simple heuristics in the analysis of language model behavior. On the one hand, if a task is solvable based on bigram statistics, it may indicate that the task may not have sufficient construct validity. Indeed, the fact that a 5-gram can score well above chance on a number of BLiMP subtasks (see Warstadt et al., 2020) could mean models do not have to learn generalized grammatical rules to solve them. On the other hand, if it can in fact support the generation of grammatical text in the majority of cases (given the pressure for shorter dependencies in natural language; see, e.g., Futrell et al., 2020), it may instead be useful to think of n-gram-like behavior as an explanation for models' observed grammatical performance rather than as a confound.

Our results provide another piece of evidence in the debate about whether learning in language models is sudden (e.g., Wei et al., 2022; Olsson et al., 2022; Power et al., 2022; Chen et al., 2024; Aoyama and Wilcox, 2025) or gradual (Schaeffer et al., 2023). In line with Kangaslahti et al. (2025), our results provide evidence that at least in some cases learning involves multiple 'hidden breakthroughs' that can underlie more apparently gradual learning trajectories seen on aggregate benchmarks. We also see that while such 'hidden breakthroughs' can lead to substantially better performance on some data subsets, they can also lead to substantially poorer performance on others, and that both of these an be relatively invisible when looking only at the aggregate score.

Taken together, our results highlight how a targeted analysis of data subsets over the course of training can provide interpretable explanations for language model behavior even without additional mechanistic analyses. With grammatical benchmarks, especially those based on comparing performance on minimal pairs (e.g., Linzen et al., 2016; Marvin and Linzen, 2018; Gulordava et al., 2018; Warstadt et al., 2020; Gauthier et al., 2020) the different versions of each sentence are interpretable by their nature. In many cases, they are drawn from—or based on—previous psycholinguistic research where these experimental manipulations are explicitly designed to highlight specific interpretable differences in human behavior. Thus, using the disaggregated versions of such datasets (as in, e.g., Arehalli and Linzen, 2020; Ryu and Lewis, 2021) allows comparisons across experimental conditions, and is thus inherently interpretable, as it is in the original human studies (see, e.g., Bock and Cutting, 1992). We move beyond such analyses at a single snapshot of model behavior by adding the dimension of time. As with humans, looking at patterns of behavior at various stages over the course of training can be informative. Specifically, we can see how performance at different conditions relative to each other varies over time, allowing us to characterize behavior as falling into different interpretable phases.

Beyond grammatical and other constructed benchmarks, such disaggregation based on *a priori* theoretical constructs is likely to be more difficult; however, such theory-driven work is likely to be key in furthering our understanding of how language models come to have the capabilities they display. Additionally, bottom-up approaches for identifying meaningfully-different subsets exist (see, e.g., Kangaslahti et al., 2025), and subsets identified this would could then be further characterized to gain a better understanding of model behavior and how it changes during training.

Limitations

This work has several limitations. First, we only investigate language model performance on English subject-verb agreement, and only consider attractors occurring within prepositional phrases. Investigating whether the trends we observe generalize beyond this limited scope would be a valuable direction for future work. Additionally, our study only uses the PolyPythia suite because we are not aware of any set of pretrained language models of multiple sizes with multiple random seeds that have a similar number of checkpoints for the early stages of training that are relevant to the present study. Finally, our work is explanatory. While this study has demonstrated the utility of disaggregation and provided a new analysis of grammatical learning in language models, it may be premature to draw any strong conclusions about the latter without further confirmatory analyses based on more specific predictions.

5 Conclusions

Our results suggest that the learning of grammatical rules such as subject-verb agreement by language models is neither sudden nor gradual—instead, it proceeds in a sequence of 'hidden breakthroughs' (Kangaslahti et al., 2025) that roughly correspond to the language model learning agreement patterns with increasingly longer dependencies. Thus, substantial insights into model behavior can be gained by analyzing performance at the level of specific subsets over the course of training and by comparing across these subsets. Indeed, our results demonstrate one can observe complex training dynamics even when analyzing behavior using a simplistic evaluation metric such as accuracy.

Acknowledgments

We would like to thank the members of the Computational Psycholinguistics Laboratory at MIT and the Language and Cognition Laboratory at UCSD for their valuable advice and discussion. James Michaelov was supported by a grant from the Andrew W. Mellon foundation (#2210-13947) during the writing of this paper.

References

Aoyama, T. and Wilcox, E. (2025). Language Models Grow Less Humanlike beyond Phase Transition. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24938–24958, Vienna, Austria. Association for Computational Linguistics.

Arehalli, S. and Linzen, T. (2020). Neural Language Models Capture Some, But Not All Agreement Attraction Effects. In Denison, S., Mack, M., Xu, Y., and Armstrong, B. C., editors, *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. cognitivesciencesociety.org.

Biderman, S., Prashanth, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. (2023). Emergent and Predictable Memorization in Large Language Models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Bock, K. and Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1):99–127.

Bock, K. and Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Chang, T. A., Tu, Z., and Bergen, B. K. (2024). Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362. Place: Cambridge, MA Publisher: MIT Press.

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. (2024). Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283–321. Place: US Publisher: American Psychological Association.

Evanson, L., Lakretz, Y., and King, J. R. (2023). Language acquisition: do children and language models follow similar learning stages? In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*. Publisher: Taylor & Francis Group.

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, 47(1):27–52. Publisher: Linguistic Society of America.

Futrell, R., Levy, R. P., and Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling.

Garrett, M. F. (1975). The Analysis of Sentence Production. In Bower, G. H., editor, *Psychology of Learning and Motivation*, volume 9, pages 133–177. Academic Press.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., and Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities. arXiv:1901.05287 [cs].

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics. container-title-short: ABBR.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Hu, J. and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., and Wilcox, E. G. (2024). Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Choshen, L., Cotterell, R., Warstadt, A., and Wilcox, E. G., editors, *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Jumelet, J., Weissweiler, L., and Bisazza, A. (2025). MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. arXiv:2504.02768 [cs].

Kangaslahti, S., Rosenfeld, E., and Saphra, N. (2025). Hidden Breakthroughs in Language Model Training. arXiv:2506.15872 [cs].

Kenney, T. J. and Wolfe, J. (1972). The acquisition of agreement in English. *Journal of Verbal Learning and Verbal Behavior*, 11(6):698–705.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Lampinen, A. (2024). Can Language Models Handle Recursively Nested Grammatical Structures? A Case Study on Comparing Models and Humans. *Computational Linguistics*, 50(4):1441–1476.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12414.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. (2024). Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. arXiv:2401.17377 [cs].

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 0(0).

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., and Clahsen, H. (1992). Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development*, 57(4):i–178. Publisher: [Society for Research in Child Development, Wiley].

Marvin, R. and Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. (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. Publisher: Proceedings of the National Academy of Sciences.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2022). In-context Learning and Induction Heads. arXiv:2209.11895 [cs].

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177 [cs].

Rumelhart, D. E., McClelland, J. L., and Group, P. R. (1986). *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

Ryu, S. H. and Lewis, R. (2021). Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage?

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A.,

Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., Melo, G. d., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

van der Wal, O., Lesci, P., Müller-Eberstein, M., Saphra, N., Schoelkopf, H., Zuidema, W., and Biderman, S. (2024). PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., and Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Wolf, T. (2019). Some additional experiments extending the tech report "Assessing BERT's Syntactic Abilities" by Yoav Goldberg. Technical report.

A Verb Frequencies

We provide the frequency of each verb form in The Pile (Gao et al., 2020), the corpus on which all language models were trained. Frequencies were calculated using the *infini-gram* web interface (Liu et al., 2024).

Table 1: Frequency of each form of each verb analyzed in The Pile.

	Singular		**Plural**	
Verb	**Word**	**Frequency**	**Word**	**Frequency**
be	is	2,055,643,528	are	816,249,141
admire	admires	97,112	admire	868,285
approve	approves	233,065	approve	1,522,021
avoid	avoids	878,590	avoid	19,190,343
confuse	confuses	159,992	confuse	637,652
criticize	criticizes	138,875	criticize	521,911
discourage	discourages	102,410	discourage	556,694
encourage	encourages	1,059,654	encourage	4,190,741
engage	engages	636,706	engage	4,491,021
greet	greets	141,839	greet	1,886,071
inspire	inspires	358,885	inspire	1,206,643
know	knows	11,077,961	know	130,967,397
observe	observes	629,563	observe	5,397,406
remember	remembers	937,143	remember	16,943,328
stimulate	stimulates	705,576	stimulate	1,522,151
understand	understands	1,385,876	understand	30,822,538

B Verb-Level Plots

We provide verb-level plots for each single-token and multi-token verb in Figure 2.

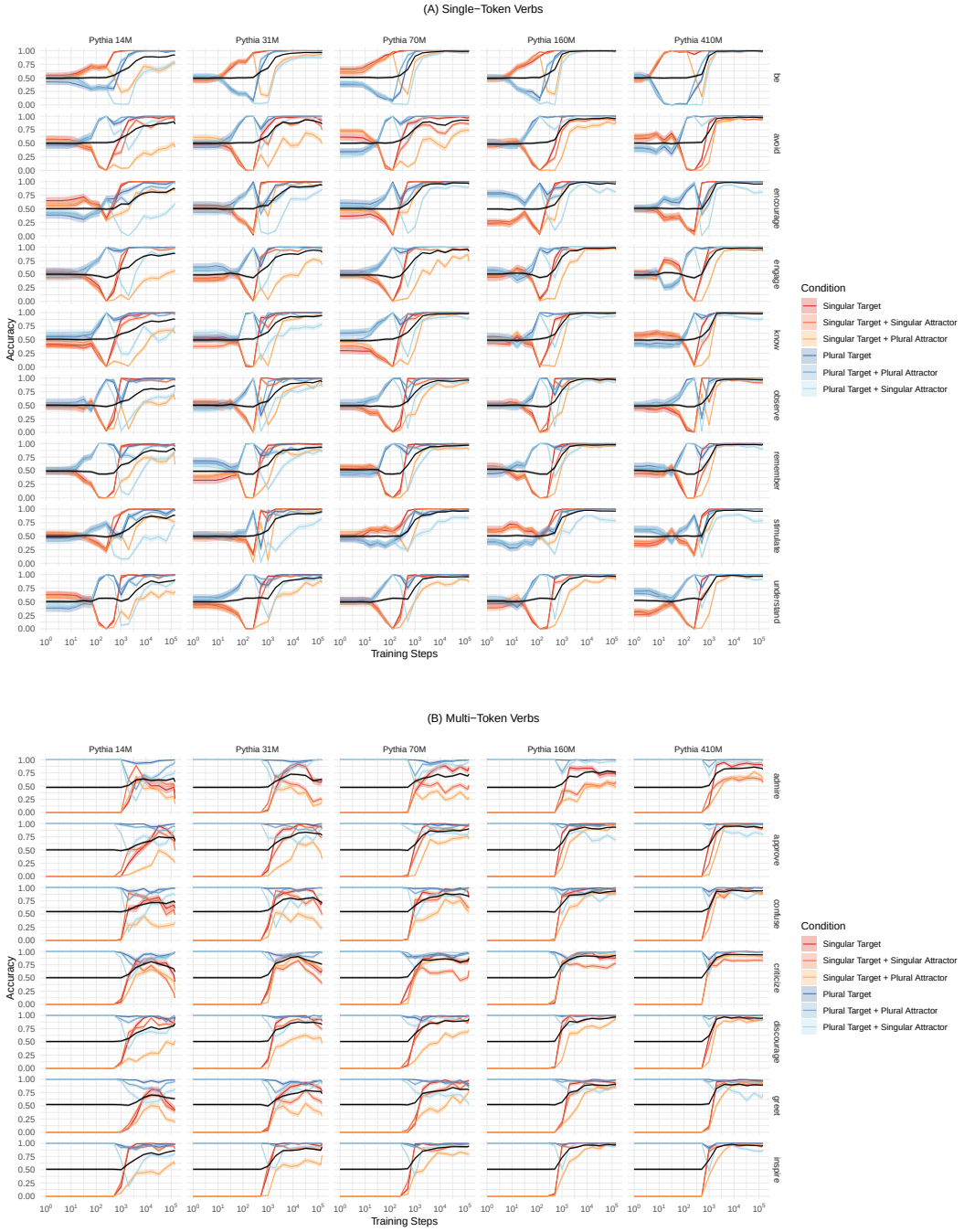

Figure 2: PolyPythia model accuracy on subject-verb agreement stimuli with (A) single-token and (B) multi-token verbs. The black line represents the mean across all conditions (i.e., the aggregate score). Shading reflects 95% confidence intervals.

C Seed-Level Plots

We provide seed-level plots for each *be* (Figure 3), other single-token verbs (Figure 4), and multi-token verbs (Figure 5).

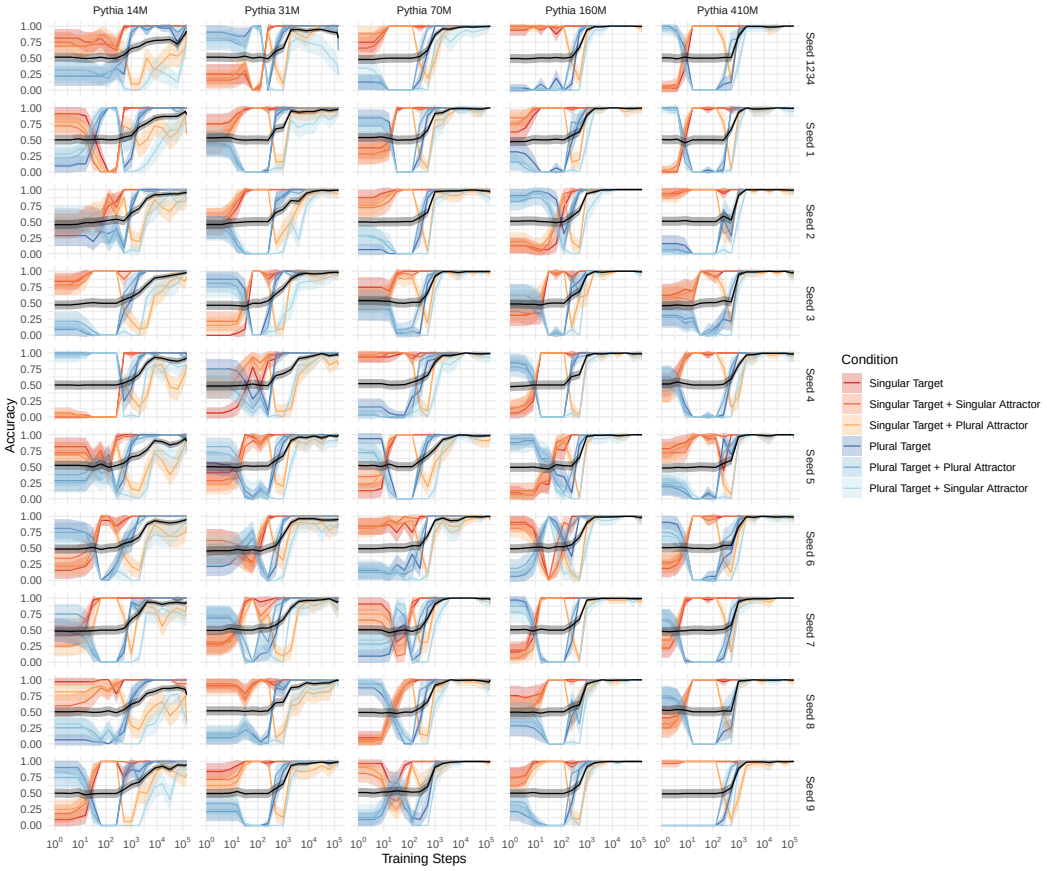

Figure 3: PolyPythia model accuracy on subject-verb agreement stimuli for the verb *be*. The black line represents the mean across all conditions (i.e., the aggregate score). Shading reflects 95% confidence intervals.

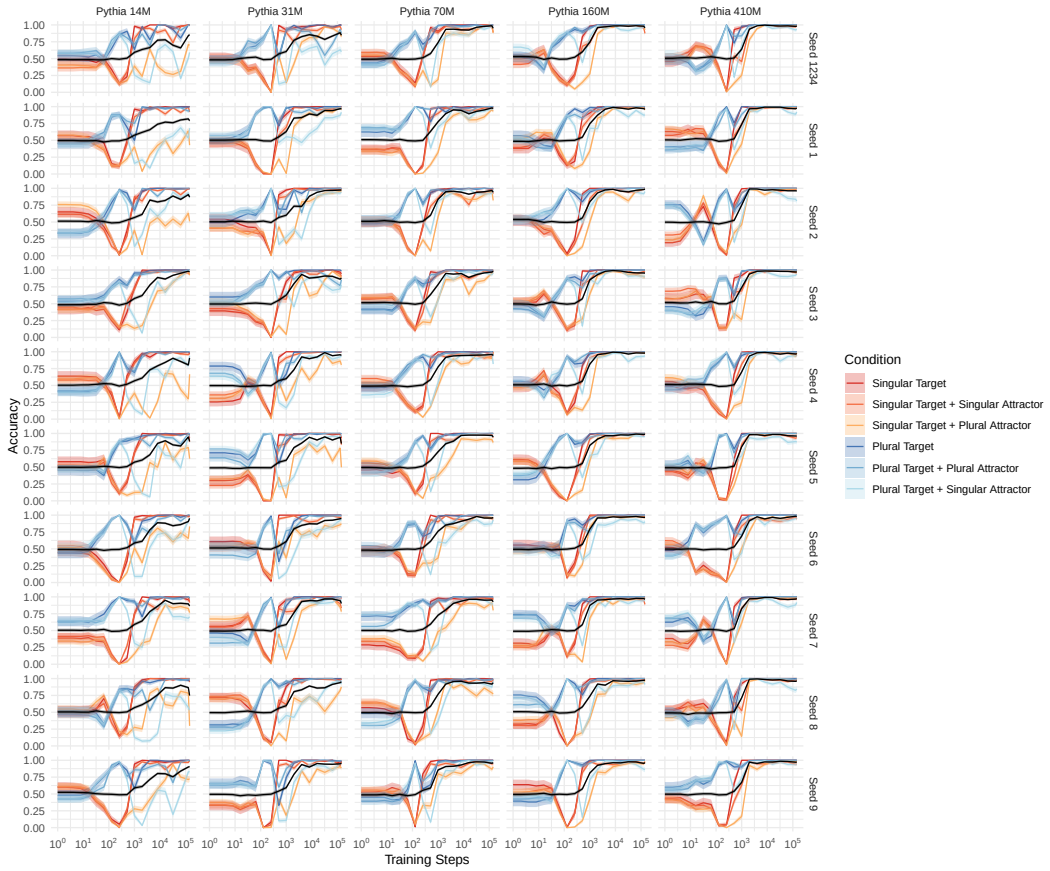

Figure 4: PolyPythia model accuracy on subject-verb agreement stimuli for single-token verbs. The black line represents the mean across all conditions (i.e., the aggregate score). Shading reflects 95% confidence intervals.

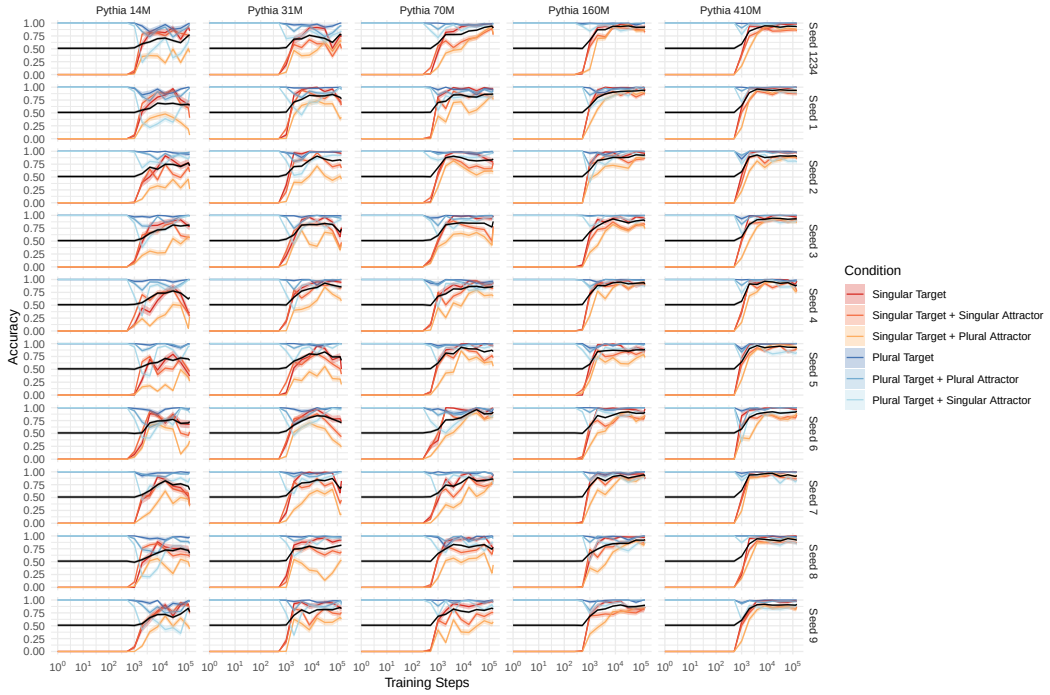

Figure 5: PolyPythia model accuracy on subject-verb agreement stimuli for multi-token verbs. The black line represents the mean across all conditions (i.e., the aggregate score). Shading reflects 95% confidence intervals.

D Token Normalization

When using language model log-probability as a way to assess their grammatical capabilities, an open question is how to score each sentence (Lau et al., 2017)—most commonly, whether to take the sum (Hu and Levy, 2023) or the mean (Jumelet et al., 2025) log-probability of each token in the relevant region (i.e., the word or sentence for which log-probability is being calculated). The former gives the 'true' log-probability assigned to the sequence by the model, while the latter normalizes to account for the fact that all else being equal, a longer token sequence will be a sum over a larger number of log-probabilities, and thus likely to have a lower log-probability in total. We carry out both analyses for the multi-token verbs (there is by definition no difference for single-token verbs). As can be seen in Figure 6, normalizing by the number of tokens (i.e., taking the mean log-probability) appears to be too strong—after the early stages of training, the mean log-probability of two-token singular verb is always higher than the log-probability of the one-token plural form.

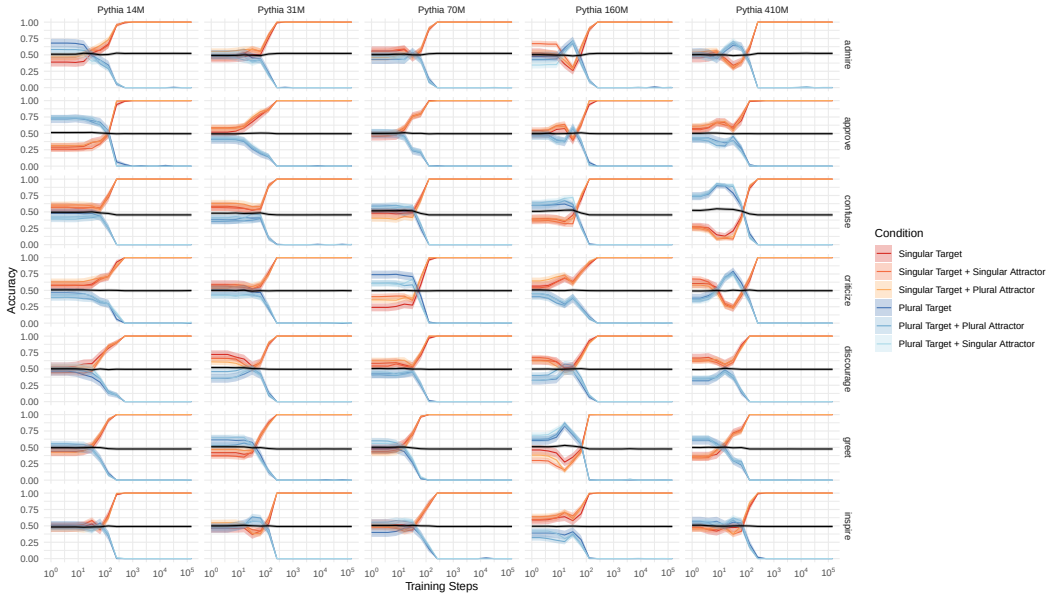

Figure 6: PolyPythia model accuracy (based on normalized log-probability) on subject-verb agreement stimuli with multi-token verbs. The black line represents the mean across all conditions (i.e., the aggregate score). The shaded areas reflect 95% confidence intervals.