Small Language Model Learning with Inconsistent Input

Anonymous ACL Submission

Abstract

Modern language models, such as GPT-3, BERT, and LLaMA, are notoriously data-hungry, requiring millions to over a trillion tokens of training data. Yet, transformer-based learning models have demonstrated a remarkable ability to learn natural languages. After sufficient training, they can consistently distinguish grammatical from ungrammatical sentences. Children as young as 14 months already have the capacity to learn abstract grammar rules from very few examples, even in the presence of non-rule-following exceptions. Yang's (2016) Tolerance Principle specifies an exact threshold for how many exceptions are permissible in a given dataset for a rule to be learnable by humans. We explore the minimal amount and quality of training data necessary for rules to be generalized by language models to test for evidence of Tolerance-Principle-like effects. We implement BabyBERTa (Huebner et al. 2021), a transformer-based language model optimized for training on smaller corpora than most LLMs. We train it on very small artificial grammar sets. For the simplest kind of rule, BabyBERTa can learn from datasets of under 1,000 tokens. The effect of type and token frequency of exemplars vs. exceptions on learning follows a gradient. We see no effect that can be related to the Tolerance Principle.

1 Introduction

1.1 Tolerance Principle

Learning a rule from a set of examples, in an unsupervised (no feedback) setting, requires the ability to generalize the rule to novel instances unseen in the training set (Huebner et al. 2021). We say that a rule is *productive* if it is learnable from a set of examples. Given a sufficient set of examples, being able to determine whether a rule should or should not be productive—i.e., having a theory that can explain how it is that abstract rules are generalized—is a challenge relevant to linguists and cognitive scientists because, among other things, a good explanation of the ability to generalize would shed a great deal of light on the process of early language acquisition in humans.

One theory of rule generalization is the Tolerance Principle (TP), originally proposed via mathematical derivation in *The Price of Linguistic Productivity* (Yang 2016). The Tolerance Principle was derived as a necessary consequence of a rule-ordering algorithm known as the Elsewhere Condition (Anderson 1969, Kiparsky 1973), which Yang proposes as a cognitive model of processing rules and exceptions. According to the Elsewhere Condition, as applied to the human brain, learning operates in an "exceptions-first, rule-later" fashion. When encountering a new exemplar and needing to decide whether to apply a rule, the brain must first consider every known exception to the rule (in order to see if this exemplar is one of these exceptions) before the general rule can be applied to it. When there are very many exceptions and very few rule-following examples, it is more time-efficient to just memorize each exemplar on a case-by-case basis and not try to learn a rule at all. The Tolerance Principle explores, mathematically, the relationship between the number of exceptions and the number of rule-following examples that allows the brain to "optimize/minimize the time complexity of language use," (Yang, 2016, p. 60).

The Tolerance Principle is well described in "A User's Guide to the Tolerance Principle," (Yang 2018), but it is made most explicit in "A User's Defense of the Tolerance Principle" with the following excerpt: "The TP is first and foremost a theory of learning. It specifies a precise threshold, as a proportion of items in the learner's experience, that a generalization can tolerate as exceptions: $\theta_N = N/lnN$, where N is the cardinality of the item set," (Yang, 2023, p. 2). Yang makes the claim that the Tolerance Principle is applicable to many kinds of learning where a rule must be generalized despite the possibility of exceptions, and is not explicitly limited to natural language rules.

One key feature of the TP is that rule learning, according to the theory, does not occur gradually; it is instead *quantal*, meaning a rule is either productive or unproductive on a given set. In other words, given a sufficient number of examples, a learner should either be able to generalize a rule, or they will be entirely unable to do so, in which case they can only memorize the behavior of the examples they were given on a case-by-case basis, with no capacity to generalize beyond them. This applies to the learning of dominant rules over an entire set, and it also applies to the learning of sub-rules for subsets of a set.

Also crucial for a full understanding of the TP is the fact that the set size N, as well as the number of permissible exceptions $e \leq \theta_N$, both refer to numbers of *unique types of items*, with no regard to the number of repetitions of items of the same type. In a linguistic context, this means the only consideration that goes into productivity is the *type* frequency (number of unique items), with no regard to the token frequency (total number of items, including repeated items of the same type). So long as a learner is exposed to enough different examples for rule learning to occur at all, the number of repetitions of individual elements of the example set will not affect the productivity of the rule.

The Tolerance Principle's application to non-human learners has not been explored.

1.2 Rule Generalization in Human Infants

A long line of research in the laboratory of Rushen Shi has investigated the abilities of human infants to generalize abstract grammar rules. Koulaguina & Shi (2013) showed that infants as young as 14 months can generalize abstract grammar rules to novel instances from relatively little training (as few as 8 exemplar sentences, repeated four times). Koulaguina & Shi (2019) showed with 14-month-olds that a training set that consisted of 50% rule-following and 50% non-rule- following sentences was insufficient for the word-order rule to be generalized, while a training set consisting of 80% rule-following and 20% non-rule-following was sufficient. They also found that it was the type frequency of the example set and not the token frequency that determined whether a word-order shift rule was productive.

Shi & Emond (2023) continued the above paradigm with more rigor, attempting to find a threshold of permissible exceptions beyond which generalizability would be impossible. They also investigated the gradual/quantal question. Their findings lent significant support to the Tolerance Principle.

1.3 Motivation

It is difficult to explain how or why 14-month-olds are so remarkably capable of generalizing abstract rules to novel instances; however, computational models are less of a black box than a human brain. When a model uses unsupervised learning to learn a rule from noisy or exception-filled data, is its learning governed by the Tolerance Principle, or something like it? This was the question that motivated our work. If it is possible to show that models can do the same thing human infants can do, examining how they do it might help explain how human infants do it.

The problem of explaining the capacity to learn is also at the forefront of language model research (see Contreras et al. 2023, Jawahar et al. 2019). Whereas there is plenty of research on how LLMs learn when they are provided with superhuman amounts of data and training, there is very limited research on their capacity to learn with small amounts of training data.

1.4 Related Studies

A few efforts have been made to optimize LLMs with the objective of achieving substantial learning from developmentally feasible quantities of training data. In the BabyLM challenge (Warstadt et al. 2023), language models were optimized to maximize learning with a training data size of 10M words or less. Huebner et al. (2021) developed the BabyBERTa transformer-based language model as a variation of RoBERTa-base (Liu 2019) and pre-trained it on as few as 5M words, simulating the input available to children aged one to six years old. Some of the best performers on the BabyLM challenge used BabyBERTa.

2 Implementation

2.1 Task

Our objective is to attempt to address the following questions: (1) What is the minimal amount of training data that our language model needs in order to learn a rule? (2) How noisy can this training data be? In other words, what proportion of training data in the training set can be non-rule-following for the rule to still be learnable? What is the relation between this proportion and the size of the dataset? (3) Is productivity quantal or gradient? That is, if a language model can generalize a rule to novel instances, is there a gradient or quantal effect as we move from unproductive regions of the parameter space to productive regions?

2.2 Model Selection

2.2.1 Architecture

We implement BabyBERTa (Huebner et al. 2021), whose code is available on GitHub. BabyBERTa uses the Transformers architecture (Vaswani 2017) and is the result of a fine-tuning of the hyper-parameters of RoBERTa (Liu 2019).

BabyBERTa, in line with RoBERTa and differing from BERT (Devlin 2018), does not do next-sentence prediction. It is instead trained only on the masked language model (MLM) pre-training objective used by BERT. A new random subsample of tokens is selected for masking every epoch.

Unlike RoBERTa-base, BabyBERTa is trained exclusively on single sentences. This means that the prediction of masked tokens takes into account only the rest of the tokens in the same sentence as the masked token. The MLM procedure is a form of self-supervised learning.

2.2.2 Hyper-Parameters

Like the original BabyBERTa implementation, our model uses 8 layers, 8 attention heads, 256 hidden units, and an intermediate size of 1024. We use Adam optimizer (Kingma 2014) with a learning rate of $1e-4$. Batch size is set to 16. In creating a random subsample of tokens for masking, tokens are selected with a probability of 0.15.

2.2.3 Training Procedure

We train our model on a text (.txt) file. The primary reason we use transformers rather than another neural network architecture is to be able to train our model on sequential text data. The simplest kind of rule, with as few features as possible, is a binary rule. We trained the model on binary strings of 0's and 1's of length 16. Our rule was: the first digit of each vector should be '1'.

In all our trials, we separated sentences in the training sets by a newline character (one vector is considered a sentence), like the original BabyBERTa's training data.

We trained the model many times from scratch, varying (a) the proportion of exceptions in the dataset, (b) the number of unique vectors (sentences) in the dataset, and, (c) the number of epochs of training.

2.2.4 Evaluation Procedure

After a full training sequence was complete, we tested the trained models on novel test sets, whose format was inspired by the grammar test suites used to evaluate BabyBERTa (Huebner et al. 2021).

Test vectors were generated in pairs. Each pair of vectors was identical, except that the first digit of the 16 digits of one of the vectors was '1'and in the other it was '0'. Each vector has its "surprisal" calculated. Surprisal is equivalent to the sum of the cross-entropy errors of each token in a given sequence. Since our sequences were only one token each, surprisal was just the cross-entropy error of that token.

If the model has learned a rule, then it should assign a lower surprisal score to a vector that follows the rule than to a nearly identical vector that breaks the rule. The model did a better job predicting one sentence in each pair over the other—the one with a lower surprisal score. We say that the model *prefers* sentences with lower surprisal scores.

The model's accuracy on each test set is equivalent to how often, as a percentage, the model prefers vectors that follow the rule, which we compute by dividing the number of vector pairs for which the model prefers the rule-following vector by the overall number of vector pairs.

For each from-scratch model, we generated a new unique test set of 1,000 vector pairs.

3 Trials

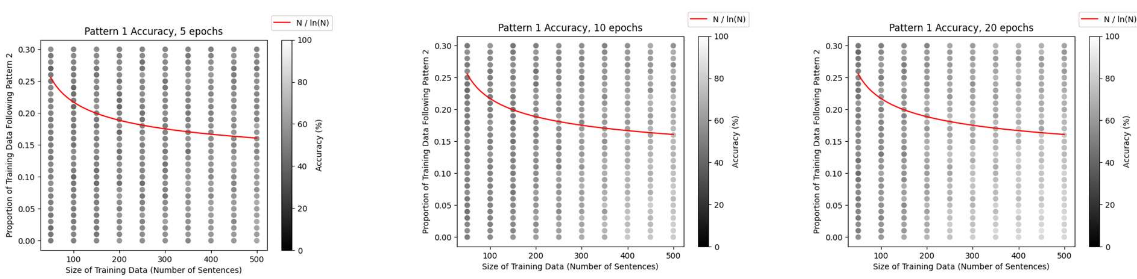
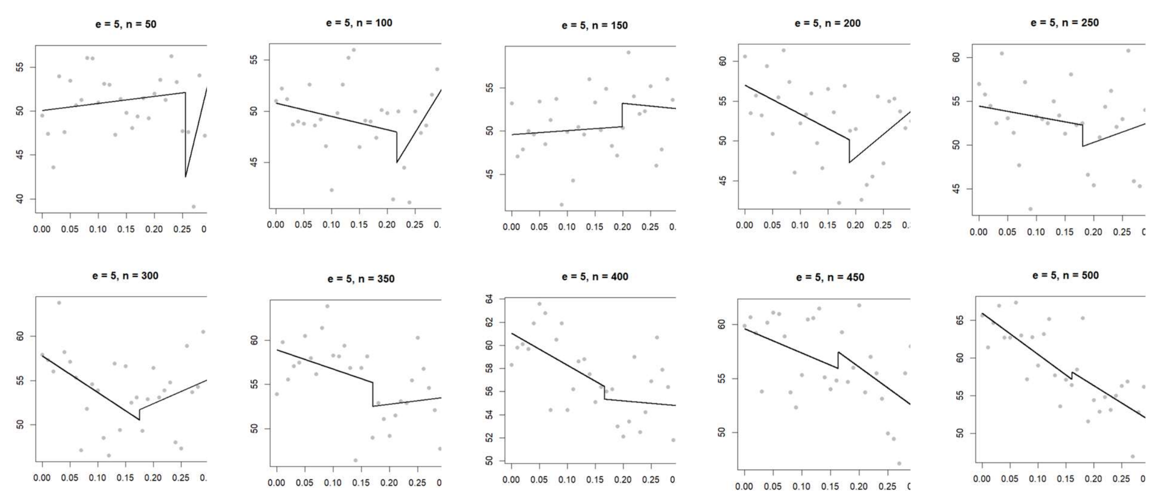


Figure 1: Model accuracies (represented by color of a point) for different training set sizes (x-axis), proportions of exceptions per training set (y-axis), and number of epochs (5, 10, & 20).



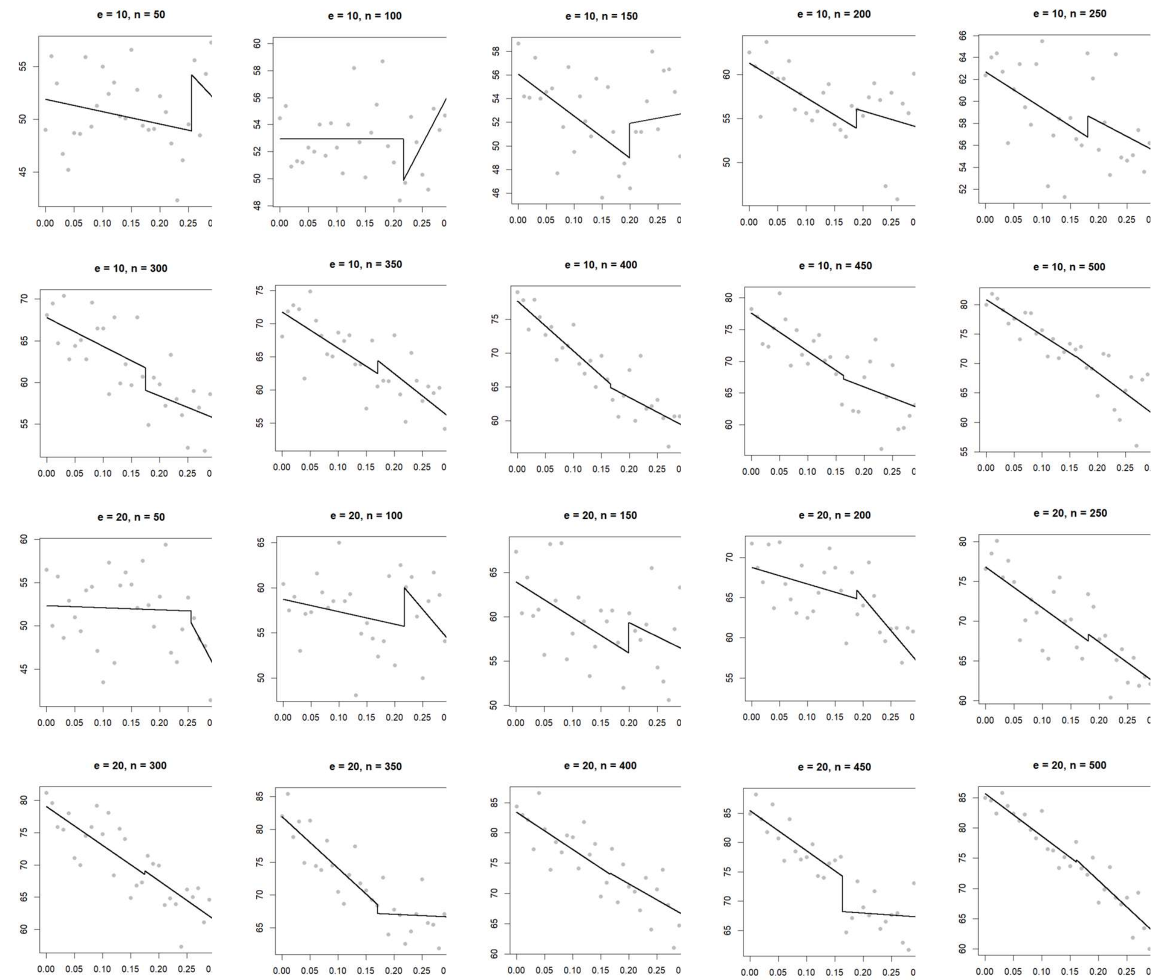


Figure 2: Effects of varying the proportion of exceptions (x-axis) on model accuracy (y-axis) for all combinations of e (# of epochs) and n (size of training dataset). Each graph contains 2 linear regressions: one on the left side of the TP threshold ($\theta_n = n/ln(n)$) and one to the right.

Figures 1 and 2 show the results of training and testing our models. If our models' learning were governed by the Tolerance Principle, we would expect:

(1) Learning should occur quantally. There should be a statistically significant jump, for each graph in Figure 2, from the regression on the left of the TP threshold to the regression on right of the TP threshold. The slope of each regression should be close to 0.

(2) Varying the number of epochs should have no significant effects on learning, since token frequency is not significant in determining whether the language model can learn.

We observe some clear trends in the above figures. For one, noticeable learning of a rule appears to be possible from training sets of just a few hundred vectors. The number of epochs in training has a major effect on learning: increasing the number of epochs leads to higher overall accuracies.

In general, we see no Tolerance-Principle-like quantal effect. In Figure 2, the jump from one regression to the other (at the TP threshold) was only statistically significant in 2 instances out of 30: (*e=5, n=50*) and (*e=20, n=450*), no more than we would expect by chance. In combinations of *e* and *n* where learning occurred, we tend to see a gradient decrease in model accuracy as the proportion of exceptions in the training set increases.

4 Conclusion

For this machine learning architecture, datasets of a few hundred examples are large enough for rule learning to occur. Learning appears to follow a gradient—as the proportion of exceptions is increased, there is a linear, not quantal, decrease in accuracy. Token frequency is significant in determining whether this language model can learn; training for more epochs over the same data increases accuracy. The threshold predicted by the Tolerance Principle seems to have no significant bearing on the language model's learning.

5 Limitations

This work does not reflect a broad study on many language models. It is limited in scope to the study of one model with fixed hyper-parameters.

References

Anderson, S. R. (1969). West Scandinavian vowel systems and the ordering of phonological rules. PhD thesis, MIT.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021, November). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624-646).

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiparsky, P. (1973). Elsewhere in phonology. In Anderson, S. R. and Kiparsky, P., editors, A festschrift for Morris Halle, pages 93–106. Holt, Rinehart and Winston, New York.

Koulaguina, E., & Shi, R. (2013). Abstract rule learning in 11-and 14-month-old infants. *Journal of psycholinguistic research*, *42*, 71-80.

Koulaguina, E., & Shi, R. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, *26*(4), 416-435.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Shi, R., & Emond, E. (2023). The threshold of rule productivity in infants. *Frontiers in Psychology*, *14*, 1251124.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*002E

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., ... & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Yang, C. (2016). *The Price of Linguistic Productivity.* Cambridge, MA: The MIT Press.

Yang, C. (2018). A user's guide to the tolerance principle. *Unpublished work*.

Yang, C. (2023). A User's defense of the tolerance principle. *University of Pennsylvania*.