

Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly

Bastian Bunzeck and Sina Zarriß

Computational Linguistics, Department of Linguistics

Bielefeld University, Germany

{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de

Abstract

Syntactic learning curves in LMs are usually reported as relatively stable and power law-shaped. By analyzing the learning curves of different LMs on various syntactic phenomena using both small self-trained llama models and larger pre-trained pythia models, we show that while many phenomena do follow typical power law curves, others exhibit S-shaped, U-shaped, or erratic patterns. Certain syntactic paradigms remain challenging even for large models, resulting in persistent preference for ungrammatical sentences. Most phenomena show similar curves for their paradigms, but the existence of diverging patterns and oscillations indicates that average curves mask important developments, underscoring the need for more detailed analyses of individual learning trajectories.

1 Introduction

The training goal of modern neural language models is simple: optimizing the prediction of the next (or a masked) token. During optimization over enormous numbers of such tokens, complex linguistic knowledge emerges as a “side effect”. But how is this knowledge and its learning trajectory characterized? Existing empirical evidence seems to suggest that morphological, syntactic and basic semantic knowledge in language models is acquired quite early during pre-training, normally with a power-law like increase over the first 5-15% of the first training epoch (*inter alia* Chiang et al., 2020; Liu et al., 2021; Saphra, 2021; Müller-Eberstein et al., 2023).

However, evaluation protocols that assess concrete learning trajectories of LMs are only beginning to emerge. Current probing approaches often mask developmental difficulties by reporting averaged scores over large and varied evaluation data sets, although, as Ritter and Schooler (2001) note, “[a]veraging can mask important aspects of

learning”. The learning curves – plots of task performance over the training period – are frequently assessed in a purely qualitative way, with little common best practices as to which training phases and how many epochs are to be described (Viering and Loog, 2023).

In this paper, we take first steps towards a more systematic analysis of the concrete learning curves for a variety of linguistic phenomena. We train a suite of small LMs, checkpointing them logarithmically during their first training epoch, test them on the BLiMP probing suite (Warstadt et al., 2020) and compare them to recent larger LMs that provide similarly-spaced checkpoints. We analyze the resulting learning curves qualitatively (in more detail than previous research), but also quantitatively (by categorizing and clustering shapes). In doing so, we are able to discern which phenomena are easier to learn and how trajectories differ between smaller and larger language models. Moreover, we investigate whether similar phenomena also exhibit similar trajectories or whether averaged learning curves obstruct some of the underlying trade-offs and instabilities of linguistic learning in LMs.

We find that when looking at the individual BLiMP paradigms and their learning curves, a more nuanced picture of how they are (not) learned emerges. While most curves do follow the prototypical power law, completely stable curves and to a lesser degree S- and U-shaped curves are also frequent. However, many paradigms also feature ill-behaved curves that never converge to stable performance or decrease over training. Inside the broader phenomenon sets, we find sheaves of curves for those mastered earlier, whereas the curves for hardly mastered phenomena exhibit strong differences. Moreover, larger models generally converge towards more power-law curves. As such, our study puts some previous results into question – certain syntactic phenomena seem to be hardly learnable even by large language models trained on massive

amounts of data, and even good performance at early training stages can deteriorate again after the model is confronted with more linguistic data.

2 Related work

Learning curves in ML and humans Every ML training run has a learning curve (target function or loss function over time), but these curves have not received much scrutiny and are often assumed to follow a power law, despite varying significantly depending on the task (Shalev-Shwartz and Ben-David, 2014; Viering and Loog, 2023). Viering and Loog (2023) review the variety of learning curve shapes, identifying both well-behaved, steadily increasing curves and ill-behaved curves that show degrading performance or oscillation. Well-behaved, monotonic curves are the common targets of ML research (Viering et al., 2019), often categorized using power law or exponential functions. In reality, not every learning curve is monotonically increasing. Exceptions include phase transitions with sudden performance boosts (Viering and Loog, 2023), peaks (Nakkiran, 2019), dips (Loog and Duin, 2012), and curves that oscillate through several maxima and plateaus (Sollich, 2001). Thus, the space of possible curve shapes is empirically much wider than often theoretically assumed.

Human learning can also be characterized by learning curves, abstracted to measurable performance on a task, with most empirical studies showing that human learning typically follows power laws (Ritter and Schooler, 2001). In language acquisition, some phenomena deviate from typical patterns. For example, past tense acquisition often follows a U-shaped curve: initially, children correctly produce high-frequency irregular and regular past forms item-based (Tomasello, 2000). As they abstract rules, they overregularize, applying regular rules to irregular verbs previously produced correctly (Saxton, 2017). Performance then gradually recovers to adult-like levels. Another common shape is the S-shaped logistic curve, where slow initial learning is followed by a rapid onset and then slow final gains. Examples are, e.g., vocabulary acquisition (Murre, 2014) or the production frequency of non-finite sentences (Hulk and Müller, 2000). However, evidence on the prevalence of and the complex trade-offs between such curves is still rather meagre. Due to a lack of empirical data, combined with small sample sizes, limited cross-

linguistic studies, and the study of very narrowly defined phenomena, some scholars argue that these effects are much weaker than assumed (e.g. Marcus et al., 1992).

Learning trajectories in LMs In their seminal paper on neural networks learning the English past tense, Rumelhart and McClelland (1986) report U-shaped learning over one epoch of training (although this development was mostly caused by their specific re-ordering of training instances, cf. Pinker and Prince, 1988). In a modern follow-up, Kirov and Cotterell (2018) find a more oscillating pattern in their LSTM-model for past-tense acquisition, although they report scores across several epochs, which hinders comparability.

Shifting the attention to the current standard in NLP, language models, it becomes apparent that investigations into learning curves are not (yet) standard practice in evaluating language models, (primarily due to the need for fine-grained checkpointing, which only few LMs provide). The more general practice of probing over time, however, is somewhat established. Chiang et al. (2020) and Liu et al. (2021) show that when comparing a variety of probing benchmarks on masked language models, syntactic information is generally acquired earlier than semantic, pragmatic and commonsense knowledge (cf. also Saphra, 2021; Teehan et al., 2022). Besides, syntactic information is also commonly located in earlier layers of LLMs (Tenney et al., 2019). Müller-Eberstein et al. (2023) analyze multiple checkpoints of the MultiBERT LMs (Sellam et al., 2022). They also find that morphological and syntactic structure is acquired very early by the models (after ~10% of the first training epoch), whereas semantic, pragmatic and general world knowledge emerge later. Their logarithmically-scaled curves still exhibit interesting, mostly S-shaped curves with a rapid take-off after a period of little learning. This is also in line with Chen et al. (2024), who find a sudden drop in training loss in masked LM training which aligns with the emergence of syntactic attention structure in attention heads.

Turning to the focus of our experiments, minimal pair tests, several additional empirical studies can be reported. Huebner et al. (2021) derive their own “Zorro” benchmark from BLiMP by excluding phenomena not found in child-directed speech. They test an extremely small (5M parameters) masked LM and show that, generally, scores improve across

	Param.	Train. tokens	Hddn. layers	Attn. heads	Embed. size	BLiMP score
baby_llama	2.97M	10M	8	8	128	64%
teenie_llama	2.97M	100M	8	8	128	67%
weenie_llama	11.44M	10M	16	16	256	67%
tweenie_llama	11.44M	100M	16	16	256	71%
pythia-14m	14M	300B	6	4	512	65%
pythia-70m	70M	300B	6	8	512	75%
pythia-160m	160M	300B	12	12	768	79%
pythia-410m	410M	300B	24	16	1024	82%
pythia-1b	1B	300B	16	8	2048	82%
pythia-1.4b	1.4B	300B	24	16	2048	82%

Table 1: Model hyperparameters of our self-trained llama models and the compared pythia models

training. They mostly show power law-like development, with the greatest improvements occurring in early stages of training. Yet, this does not apply to all included phenomena – some are never learned well (e.g. island effects or anaphor agreement). These show diminishing accuracy after early performance peaks – a fact not further discussed.

Liu et al. (2021) also examine BLiMP development during the training of a masked LM and find that their curves, which categorize phenomena more coarsely, converge to stable performance quickly, approximating power-law curves after about 20% of pre-training. Morphological and short-distance syntactic phenomena are mastered fastest, while more complex syntactic aspects, like island effects, take longer. This pattern holds for other linguistic probes, but benchmarks testing common sense or reasoning exhibit unstable behavior with oscillating curves and performance dips. Choshen et al. (2022) take a similar approach with autoregressive LMs (GPT-2, TransformerXL). They find that grammatical phenomena are acquired in a stable order along classical linguistic layers. However, not all curves show monotonic improvement; some syntax and morphology paradigms never reach stable performance and deteriorate over training. This behavior is consistent across different initializations of both architectures but does not apply to phenomena involving semantic knowledge.

3 Methods

3.1 Investigated models

We analyze two different model architectures, four self-trained llama models (Touvron et al., 2023a) and six models from the pythia family (Biderman et al., 2023).

Data We train our models on the BabyLM data set (Warstadt et al., 2023). It features written and spoken source corpora that span a wide range of registers – child-directed speech/text, adult conversations, movie dialogue, and data from Wikipedia and Project Gutenberg. Before training our models, we clean the data from artefacts, adapting scripts by Timiryasov and Tastet (2023). The pythia models are trained on The Pile (Gao et al., 2020), a 300B token corpus sourced from the internet, academic literature, code from GitHub and, to a lesser degree, spoken language, which makes it more comparable to regular LLM training corpora.

Models and training hyperparameters We use the transformers library (Wolf et al., 2020) to train four different llama models¹ (Touvron et al., 2023b). Our smallest model we call baby_llama. The larger models are differentiated by more tokens (100M instead of 10M for teenie_llama), more weights (11.44M instead of 2.97M for weenie_llama) or both in the case of tweenie_llama. As training hyperparameters, we chose a batch size of 16, 200 warmup steps, and a learning rate set to 3e-4 in accordance with Touvron et al. (2023a). From the pythia suite of GPT-NeoX models (Andonian et al., 2023), we take the six smallest models, ranging from 14M to 1.4B parameters. They were all trained on the same data, but with different model hyperparameters (cf. Table 1). Our models were trained on a single NVIDIA RTX A4000 GPU, contrasting with the pythia models trained on clusters of 32–64 GPUs.

3.2 BLiMP performance

We test BLiMP performance with lm-eval-harness (Gao et al., 2022). By calculating perplexity for the sentences in each

¹Available at <https://huggingface.co/bbunzeck>










	Shape	Graphical	Description
Well-behaved	U		Medium performance followed by a dip, then rapid improvement and stabilization
	S		Initially no learning, then rapid onset and finally stabilization
	Pow		Rapid early learning, followed by stabilization and no further gains
	Stable		No change in performance across training (standard deviation < 0.2)
Ill-behaved	InvU		Inverse U-shape, stabilization after a performance peak and subsequent decrease
	RevU		Dip in performance, stabilization on lower level than before dip
	RevS		Reversed S-curve, early performance is good, but then diminishes rapidly and never recovers
	RevPow		Reverse power-relationship – performance degradation at end of training
	Osc		Performance never stabilizes and jumps between better and worse scores

Table 2: Overview of proposed curve shapes

pair, BLiMP can be used to discern whether a grammatical sentence is preferred by an LM (less perplex): an accuracy of 50% equals the random baseline. BLiMP covers 12 different linguistic phenomena from morphology, syntax and semantics (or their interfaces), with 67 included paradigms (individual data sets). We deliberately chose BLiMP due to its widespread use and the wealth of previous results, although it suffers from problems like semantically implausible sentences (see (Vazquez Martinez et al., 2023) for more criticism and alternative data sets).

3.3 Analyzing learning curves

In line with Viering and Loog (2023), our learning curves are based on performance changes over precisely one training epoch. This choice allows us to capture the learning potential from the data upon initial exposure and observe trajectories as models encounter new data continuously. Recognizing that many linguistic phenomena are acquired early in training, we look at logarithmically spaced evaluation checkpoints: 10 checkpoints within the first 10% of training and 9 additional checkpoints until the epoch’s completion.

Assessing the shapes of learning curves systematically is a complex task. We qualitatively assign shapes, aided by fitting fifth-degree polynomials to each curve. Our categorization includes well-behaved curves, such as S-shaped, U-shaped, and power law curves, as observed in the acquisition literature. We also identify ill-behaved curves by their inverted (mirrored on the x-axis) and reversed (mirrored on the y-axis) variants. Additionally, curves that remain stable from the earliest training steps (standard deviation < 0.02) are considered well-behaved, whereas curves that oscillate continuously without stabilizing are deemed ill-behaved. A summary of our systematization is provided in Table 2.

To further examine similarities between mod-

els and paradigms, we define a feature vector for each model-paradigm combination by computing the performance differences between all successive pairs of training steps across all BLiMP paradigms. This allows us to represent each model-paradigm combination as a point in a high-dimensional vector space.

4 Results

BLiMP After one training epoch, our baby_llama achieves a general BLiMP accuracy of 64%, improving to 67% with more data (teenie_llama) or more parameters (weenie_llama). The combination of both (tweenie_llama) reaches 71%. The smallest pythia model (14M parameters), despite being trained on much larger datasets, only achieves 65%, increasing to 75% for the 70M model and 79% for the 160M model. The largest pythias (410M, 1B, 1.4B) all reach 82%, close to peak BLiMP performance reported in the original BLiMP paper (83% by GPT-2), the highest score on the HELM evaluation database (84%, Liang et al., 2023), and the best BabyLM model (86%, Warstadt et al., 2023). Therefore, our results are comparable to even larger models. As an ablation, an untrained llama model performed similarly to the random baseline, scoring 51%.

Variation in phenomenon-averaged curves In the spirit of earlier analyses, we first consider the averaged curve shapes (over the phenomenon sets in BLiMP) from a qualitative viewpoint (see Figure 1, which contrasts the smallest and largest models investigated). For the smallest llama model, the learning curves exhibit a range of shapes, including power-law curves, S-shaped curves and U-shaped curves. Many curves do not show any improvements over the training epoch. The first 10% of training is marked by the highest degree of variation, but many performance gains also happen later than that. Here it already becomes apparent that

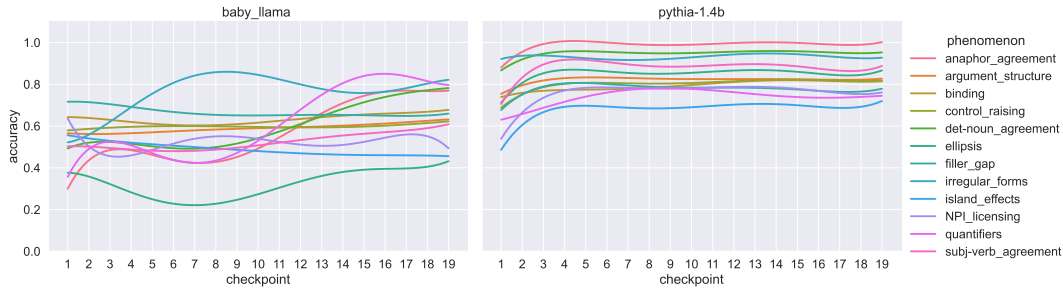


Figure 1: Learning curves over one epoch for the smallest llama and largest pythia model, averaged across BLiMP phenomena (for both models, the first ten checkpoints correspond to the first 10% of training, the following nine to remaining 90%)

more training on natural language data does not linearly improve performance on linguistic probing tasks. The largest pythia model, in contrast, displays learning curves that mostly resemble power law curves. Improvements are concentrated within the first 4-5% of training, after which the performance remains relatively stable across all phenomena (although minor performance tops and dips are observable).

We present a more detailed visualization of individual curves, categorized by phenomenon and model, in Figure 2.

Individual curves are frequently ill-behaved

The first striking observation is found in the many ill-behaved curves. For the llama models, more than a quarter of the learning curves are ill-behaved, while for the pythia models, this is true for more than one fifth (distributions found in Table 3). While larger models generally do distinguish more minimal pair paradigms effectively, those phenomena that exhibit unstable and erratic curves in smaller models frequently continue to do so in larger models. Apart from that, smaller models have a higher number of curves that remain well below 50%, indicating that for some phenomena, these models actively prefer the ungrammatical variant. This issue occurs only sporadically in larger models, for example with selected paradigms concerning quantifiers or filler-gap phenomena.

Patterns inside phenomenon sets: sheaves and divergence

Another striking aspect visible in Figure 2 is that sheaves of curves – curve sets that are close across training and show very similar shapes – are found across all models for different phenomena (e.g. argument structure and determiner-noun agreement). They become more power-law-like as the models increase in size. Apart from sheaves,

we can also find diverging patterns, where some curves inside one phenomenon show strong improvements and other curves exhibit deteriorating performance over one training epoch, for example with subject-verb agreement and filler-gap phenomena. Such diverging patterns are more prevalent in smaller models, where they often appear as almost mirrored curves. In larger models, divergent patterns are less pronounced, but for phenomena prone to divergence, some curves still tend to worsen in the largest models.

The effects of model and data size The relationship between model size and performance is not straightforward. Our llama models scale in both parameters and dataset size, while the pythia models only scale in parameters but are trained on the same number of significantly more tokens. This increased amount of data results in less granularity in our analysis. The smallest pythia model, with few parameters but a large amount of training data, still exhibits many ill-behaved curves. Also, its curves show a later take-off in BLiMP performance compared to its larger sister models.

	llama models	pythia models
Ill-behaved	27.24%	22.39%
Power law	33.21%	45.77%
S-shaped	12.32%	13.18%
Stable	14.93%	14.67%
U-shaped	12.32%	3.98%

Table 3: Percentage of curve types for both model families

When quantitatively comparing the distributions of curves among all investigated models, a clear division emerges between llama and pythia models. This is visualized in Figure 3. The division is particularly pronounced in larger pythia models with

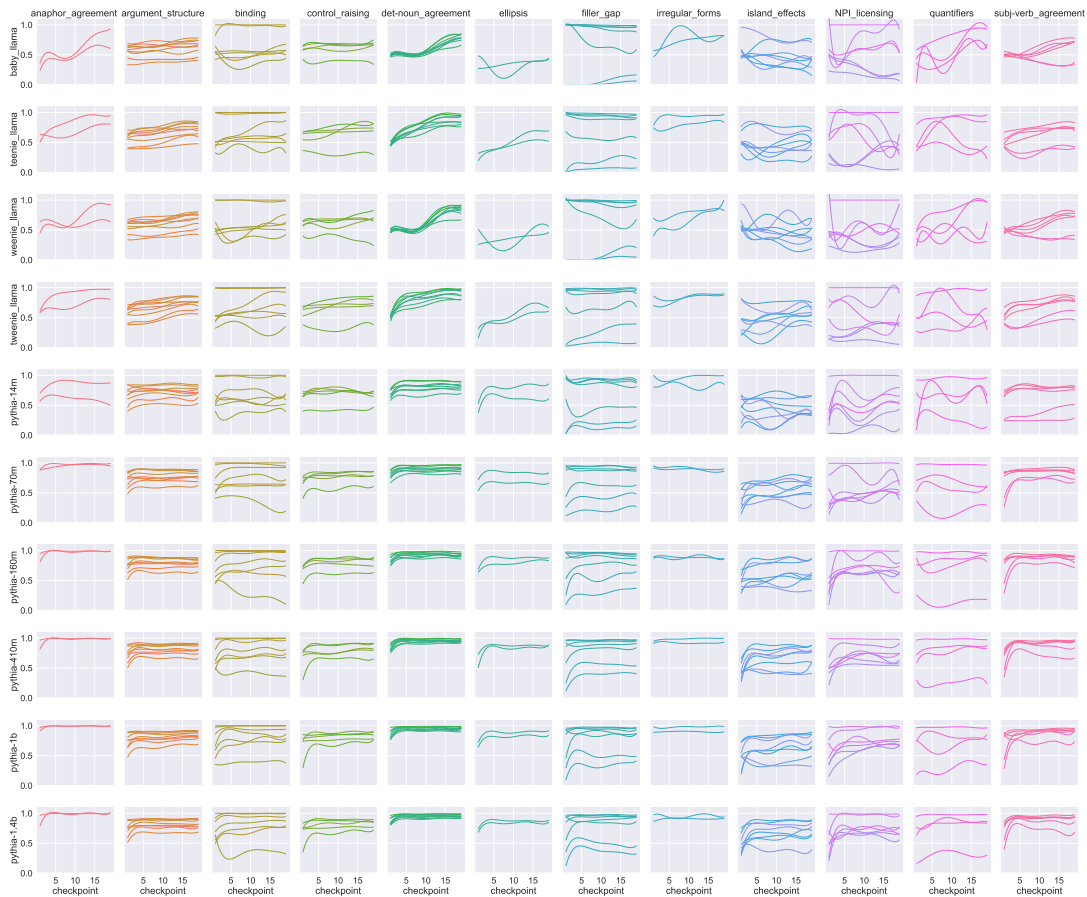


Figure 2: Learning curves for all paradigms in BLiMP, separated for models (rows) and phenomenon sets (columns)

160 million or more parameters. Llama models and the two smaller pythia models have a higher proportion of U-shaped and S-shaped curves. These models also display more ill-shaped patterns, ranging from 25% to 40%. Llama models trained on larger datasets have over 40% power-law curves, whereas the smallest pythia model shows no power-law curves. Larger pythia models have over half of their learning curves following a prototypical power-law pattern, with a significant number of stable curves (20-25%), few U-shaped developments, and no S-shaped developments. Additionally, they exhibit fewer ill-behaved curves (15-20%) compared to smaller models. The four largest pythia models show very little variation, further highlighting this distinction.

Clustering training trajectories To assess further commonalities between paradigms or models, we visualize the developmental trajectory vectors, reduced in dimensionality using t-SNE (van der Maaten and Hinton, 2008), in a scatter plot (Figure 4) and visually examine whether they form specific clusters. Initially, the plot presents a

messy picture with little visible structure. Clustering effects for different models or model architectures appear rather weak. However, clustering effects are more pronounced for BLiMP phenomena. We observe clusters for argument structure, determiner-noun agreement, and subject-verb agreement—phenomena that typically form sheaves. Additionally, NPI licensing, binding, and filler-gap phenomena also cluster, even though their curve shapes are quite varied. Conversely, there are no discernible patterns for phenomena like quantifiers or irregular forms.

Turning points across training The diverging mirrored curves described earlier in Section 4 also indicate another pattern: the minima for many paradigms coincide with the maxima for others. The point plots in Appendix B show checkpoint-wise deviations from mean performance, revealing particularly strong positive and negative deviations at certain checkpoints. The effects are especially pronounced in Figure 10, where almost all NPI paradigms show their maximum performance in the last few checkpoints, except for

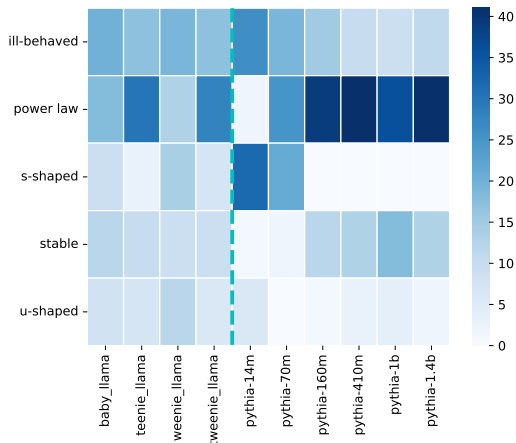


Figure 3: Co-occurrence frequencies between models and curve shapes (darker rectangles indicate higher frequency), the dashed line separates llama and pythia models

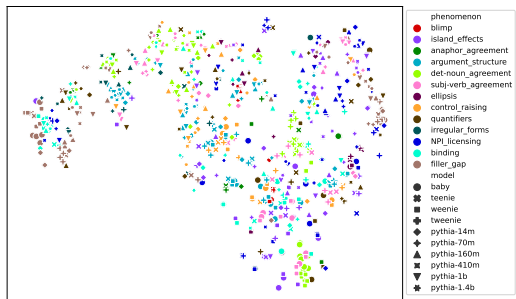


Figure 4: Dimensionality-reduced scatterplot of curve development for each model-paradigm combination

only_npi_licensor_present, which has deteriorated from earlier maxima in the first 10% of training

Key results From our qualitative and quantitative analyses, the most striking observations can be summarized as follows:

- Ill-behaved curves occur across all models, though they are less frequent in larger models with more internal parameters. When looking at non-averaged curves, these ill-behaved developments are much more pronounced-
- For many phenomenon-model combinations, the curves for related paradigms emerge as similarly shaped sheaves of individual curves. This is particularly true for, e.g., argument structure or determiner-noun agreement.
- In contrast to the aforementioned sheaves also diverging patterns are observed within phenomena. Some paradigms within the same

phenomenon have mirrored learning trajectories, where improvement in one paradigm is directly correlated with diminishing performance in another. This divergence is particularly pronounced for filler-gap phenomena, as well as in subject-verb agreement and binding.

- Shape-wise similarities are more pronounced for phenomena across different models, whereas (especially for the smaller models) there is high variation within models.

5 Discussion

Our results indicate that larger models perform better, exhibiting higher BLiMP scores, fewer ill-behaved curves, and more power-law curves, aligning with existing literature on scaling dynamics (Warstadt et al., 2020, 2023). In our self-trained llama models, improvements are seen both with increased parameters and more data, with the combination leading to even greater enhancement. Interestingly, the smallest pythia model, despite being trained on significantly more tokens compared to the llama models, performs worse on BLiMP. This suggests that in the very small pythia model, the real learning of linguistic features only begins after a large number of tokens are seen, whereas in our smaller llama models, this learning occurs much earlier. A possible explanation for this discrepancy could be the higher quality of the datasets used to train our llama models (BabyLM 10M and 100M), which offer a wider variety of genres and registers, compared to the web-sourced “The Pile” dataset used for all pythia models.

Our findings largely confirm, but also revise and expand upon, earlier reports of rapid syntax learning in language models. Many phenomena are acquired quickly (as in Liu et al., 2021, also Müller-Eberstein et al., 2023), yet some BLiMP paradigms are never fully mastered as in Huebner et al., 2021 for Zorro or Choshen et al., 2022 for BLiMP). The learning trajectories are non-linear; more tokens do not necessarily improve performance. Phenomena exhibit various curve shapes – some start strong, dip, and stabilize, while others oscillate indefinitely. Even in the largest models, certain phenomena remain unlearned, showing a persistent preference for ungrammatical sentences. This aligns with literature identifying these phenomena as difficult to learn, often displaying unusual learning curve patterns. Easily learned phenomena have organized sheaves of curves, while hard-to-learn phenomena

exhibit scattered individual curves, suggesting that phenomena based on similar linguistic features are not uniformly grounded in the same ML features.

Some peculiarities found in our analyses might be caused by BLiMP itself. For example, the `principle_A_case_1` paradigm, which exhibits almost only stable and perfect learning curves, always features a possessive pronoun (e.g. *her*) in the grammatical sentence and a reflexive (e.g. *herself*) in its ungrammatical counterpart. However, possessives are much more frequent in language than reflexives (e.g. *her*: 1.517.948 tokens vs. *herself* 56.741 tokens in ukWaC, Baroni et al., 2009), so it is reasonable to assume that a sentence containing a reflexive always has a higher perplexity. For a randomly shuffled training corpus that is representative and balanced (in the sense of Stefanowitsch, 2020, 28), these patterns should be learned very quickly from little data and thus have such a stable learning curve, whereas other phenomena that are less tied to frequency differences might not use such easy surface heuristics. Similar criticisms, e.g. about problems with the quality of example sentences, have been put forward by, *inter alia*, Vazquez Martinez et al. (2023).

An ML-based explanation for such peculiarities is that models pick up orthogonal features – features that improve performance on some paradigms within a phenomenon but degrade performance on others – during the learning process (Choshen et al., 2022). It remains open whether ML features must necessarily correspond to those considered important in linguistic theory. The presence of mirrored curves/turning points also supports the hypothesis of orthogonal features.

Finally, BLiMP’s choice of target phenomena is heavily influenced by generative, syntax-centric linguistics. Other contemporary linguistic theories (e.g. usage-based linguistics, construction grammar) might not find these phenomena particularly meaningful. In construction grammar, argument structure is determined by constructional patterns, allowing verbs to take new arguments and convey new meanings (Goldberg, 2013). Therefore, perfect performance on BLiMP may not necessarily be a desirable goal, as it might not reflect the flexible and creative language use characteristic of humans. Additionally, grammaticality is a contested notion, difficult to measure, often gradient, and strongly influenced by socio-cultural factors (Vogel, 2018, 2019). Consequently, stable curves might only be desirable for phenomena that exhibit less gradience

in human evaluation, whereas worse scores and eternal oscillations might entail better linguistic generalizations for less clear-cut paradigms.

6 Conclusion

Our study set out to characterize linguistic learning in language models through an analysis of learning curves. We conclude that while the rapid syntax learning assumption from earlier studies generally holds, it also needs revision. When averaging across many phenomena and paradigms, performance gains appear to follow a prototypical power law. However, this is not true when examining individual phenomena, many of which exhibit ill-behaved curves. Stability in BLiMP performance is often an illusion; stable average curves are based on oscillating and heavily changing minimal pair paradigms within them. With larger models and more data, there is a general shift towards greater stability and more power law curves, but even in very large models, not everything works perfectly.

On a meta-level, our study demonstrates that analyzing learning curves is a powerful tool for better characterizing learning processes. Many benchmarks include systematically organized sub-phenomena, and our methodology can illuminate specific performance developments and complex trade-offs during the learning process. This highlights the need for the community to develop best practices for reporting learning curves, categorizing their shapes, and determining the appropriate granularity for analysis across one or several epochs. Researchers should be cautious with their interpretations, as the complexity and variety of learning curves suggest a more nuanced approach is necessary.

Future work could expand on our findings by exploring how controlling for distributions of linguistic data, like Wei et al. (2021) describe, changes the curves and learning success, which would further enhance our understanding of language model learning dynamics in a more restricted setting.

Acknowledgements

We thank Simeon Junker and Henrik Voigt for their helpful comments on earlier drafts of this paper.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A02.

References

- Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. [GPT-NeoX: Large scale autoregressive language modeling in PyTorch](#).
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. [The WaCky wide web: A collection of very large linguistically processed web-crawled corpora](#). *Language Resources and Evaluation*, 43(3):209–226.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling](#). *Preprint*, arxiv:2304.01373.
- Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.
- Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. [Pretrained Language Model Embryology: The Birth of ALBERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.
- Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. [The Grammar-Learning Trajectories of Neural Language Models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#). *Preprint*, arxiv:2101.00027.
- Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Fazz, Niklas Muennighoff, Thomas Wang, Sdtblck, Ttyuntian, Researcher2, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, Kkawamu1, Xagi-Dev, Uyhcire, Andy Zou, Ben Wang, Jordan Clive, Igor0, Kevin Wang, Nicholas Kross, Fabrizio Milo, and Silentv0x. 2022. [EleutherAI/Im-evaluation-harness: V0.3.0](#). Zenodo.
- Adele E. Goldberg. 2013. [Argument Structure Constructions versus Lexical Rules or Derivational Verb Templates](#). *Mind & Language*, 28(4):435–465.
- Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. [BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.
- Aafke Hulk and Natascha Müller. 2000. [Bilingual first language acquisition at the interface between syntax and pragmatics](#). *Bilingualism: Language and Cognition*, 3(3):227–244.
- Christo Kirov and Ryan Cotterell. 2018. [Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince \(1988\) and the Past Tense Debate](#). *Transactions of the Association for Computational Linguistics*, 6:651–665.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihito Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekogonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. [Holistic Evaluation of Language Models](#). *Preprint*, arxiv:2211.09110.
- Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. [Probing Across Time: What Does RoBERTa Know and When?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Marco Loog and Robert P. W. Duin. 2012. [The Dipping Phenomenon](#). In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Georgy Gimel’farb, Edwin Hancock, Atsushi Imiya, Arjan Kuijper, Mineichi Kudo, Shinichiro Omachi, Terry Windeatt, and Keiji Yamada, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626, pages 310–317. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Gary F. Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harold Clahsen. 1992. [Overregularization in Language](#)

- Acquisition. Monographs of the Society for Research in Child Development*, 57(4):i.
- Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. [Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.
- Jaap M. J. Murre. 2014. [S-shaped learning curves](#). *Psychonomic Bulletin & Review*, 21(2):344–356.
- Preetum Nakkiran. 2019. [More Data Can Hurt for Linear Regression: Sample-wise Double Descent](#). *Preprint*, arxiv:1912.07242.
- Steven Pinker and Alan Prince. 1988. [On language and connectionism: Analysis of a parallel distributed processing model of language acquisition](#). *Cognition*, 28(1-2):73–193.
- F.E. Ritter and L.J. Schooler. 2001. [Learning Curve, The](#). In *International Encyclopedia of the Social & Behavioral Sciences*, pages 8602–8605. Elsevier.
- David E. Rumelhart and James L. McClelland. 1986. [On Learning the Past Tenses of English Verbs](#). In *Parallel Distributed Processing*, volume 2, pages 535–551. MIT Press, Cambridge, MA.
- Naomi Saphra. 2021. [Training dynamics of neural language models](#).
- Matthew Saxton. 2017. *Child Language: Acquisition and Development*, 2nd edition edition. SAGE, Los Angeles.
- Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D’Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2022. [The MultiBERTs: BERT Reproductions for Robustness Analysis](#). *Preprint*, arxiv:2106.16163.
- Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*, 1 edition. Cambridge University Press.
- Peter Sollich. 2001. [Gaussian process regression with mismatched models](#). In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.
- Anatol Stefanowitsch. 2020. *Corpus Linguistics: A Guide to the Methodology*. Language Science Press, Berlin.
- Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. [Emergent Structures and Training Dynamics in Large Language Models](#). In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 146–159, virtual+Dublin. Association for Computational Linguistics.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT Rediscovered the Classical NLP Pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Inar Timiryasov and Jean-Loup Tastet. 2023. [Baby Llama: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 251–261, Singapore. Association for Computational Linguistics.
- Michael Tomasello. 2000. [The item-based nature of children’s early syntactic development](#). *Trends in Cognitive Sciences*, 4(4).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [LLaMA: Open and Efficient Foundation Language Models](#). *Preprint*, arxiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [LLaMA 2: Open Foundation and Fine-Tuned Chat Models](#). *Preprint*, arxiv:2307.09288.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Hector Javier Vazquez Martinez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. [Evaluating Neural Language Models as Cognitive Models of Language Acquisition](#). In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalization in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

- Tom Viering and Marco Loog. 2023. [The Shape of Learning Curves: A Review](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819.
- Tom Viering, Alexander Mey, and Marco Loog. 2019. Open problem: Monotonicity of learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3198–3201. PMLR.
- Ralf Vogel. 2018. Sociocultural determinants of grammatical taboos in German. In Liudmila Liashchova, editor, *The Explicit and the Implicit in Language and Speech*, pages 116–153. Cambridge Scholars Publishing.
- Ralf Vogel. 2019. [Grammatical taboos: An investigation on the impact of prescription in acceptability judgement experiments](#). *Zeitschrift für Sprachwissenschaft*, 38(1):37–79.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 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*, pages 1–6, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The Benchmark of Linguistic Minimal Pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. [Frequency Effects on Syntactic Rule Learning in Transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Learning curves for all paradigms

In consideration of legibility and brevity, detailed plots in the appendix are provided as downsized vector graphics. Interested readers may zoom in for finer detail and further examination.

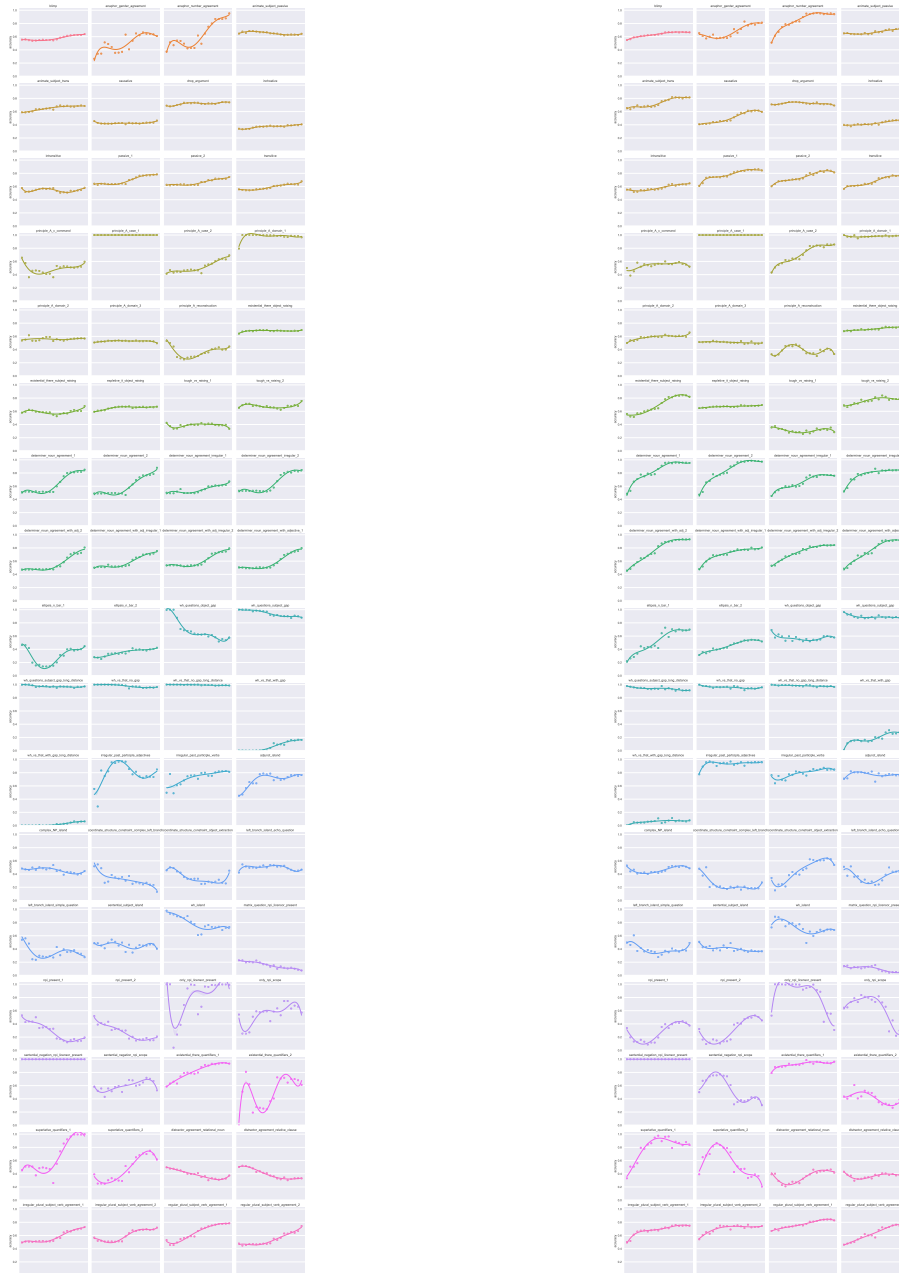


Figure 5: Learning curves for baby_llama and teenie_llama

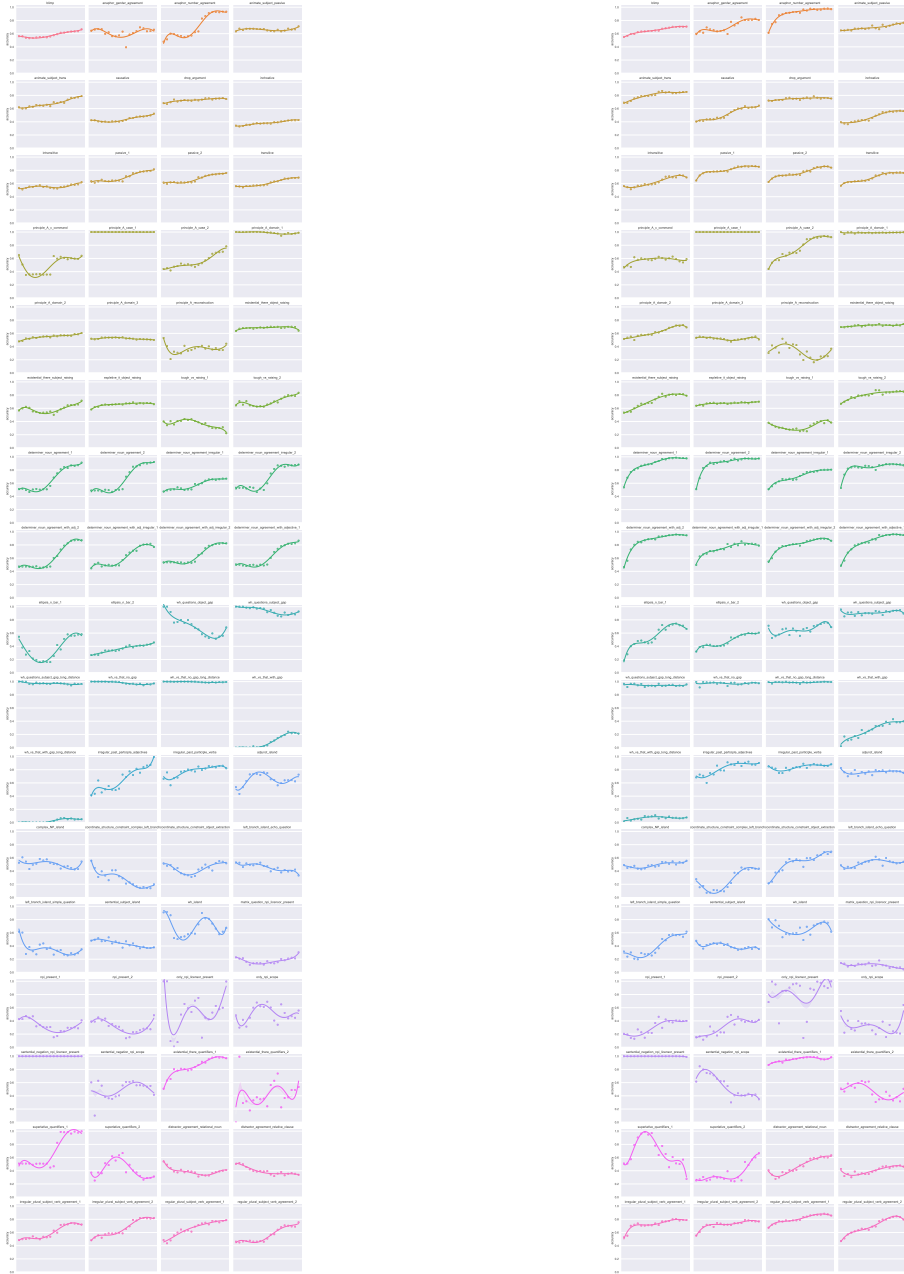


Figure 6: Learning curves for weenie_llama and tweenie_llama

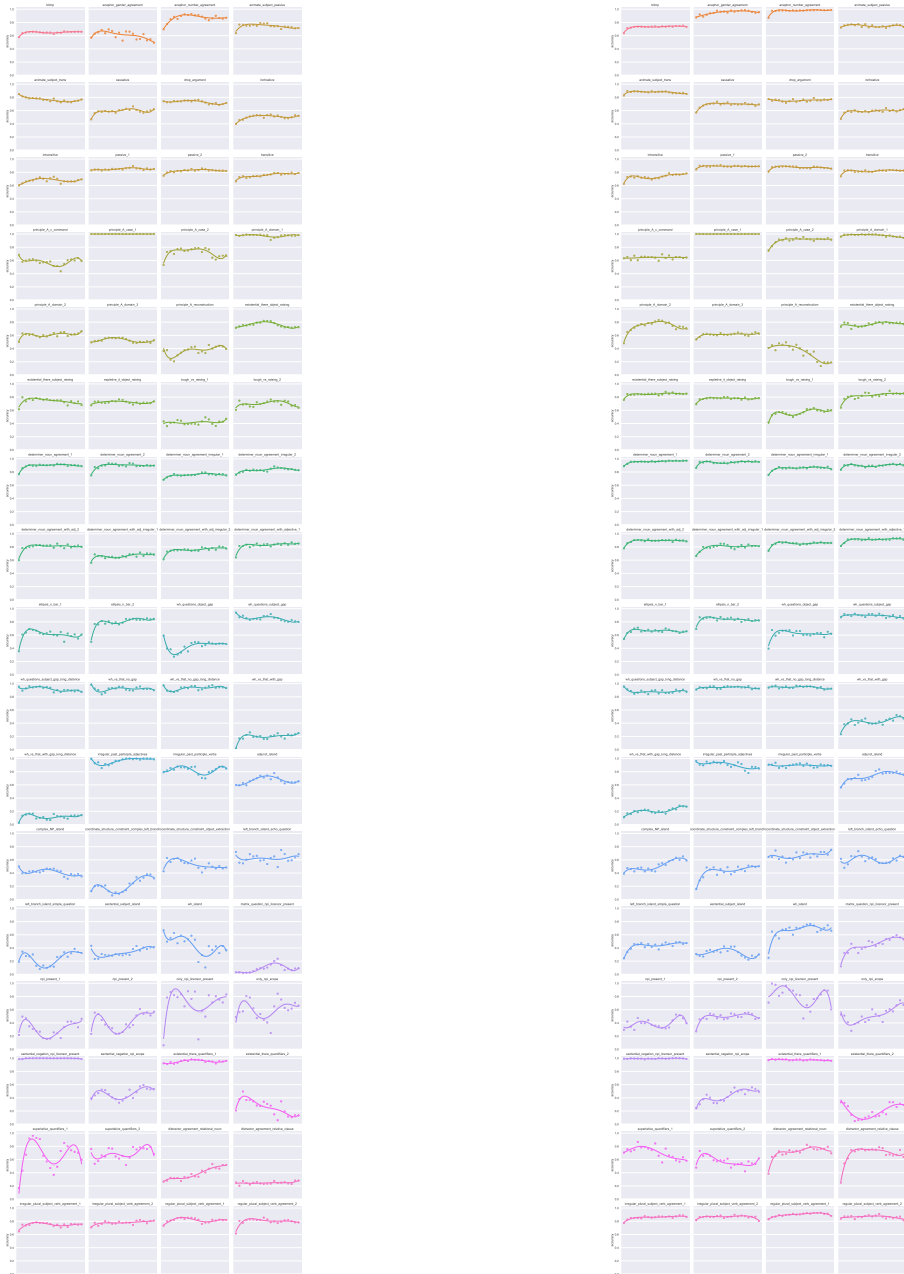


Figure 7: Learning curves for pythia-14m and pythia-70m

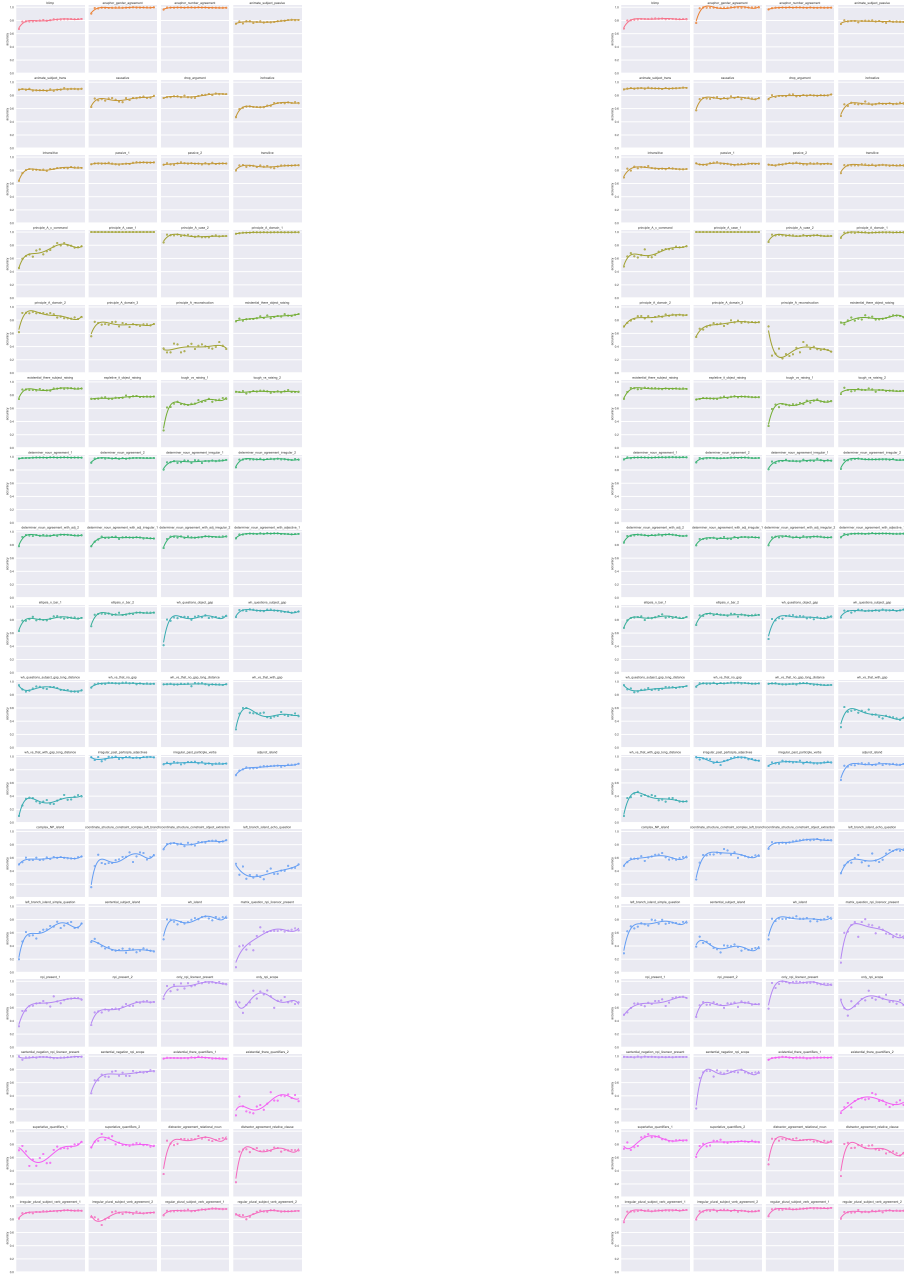


Figure 9: Learning curves for pythia-1b and pythia-1.4b

B Point plots for distance to mean performance

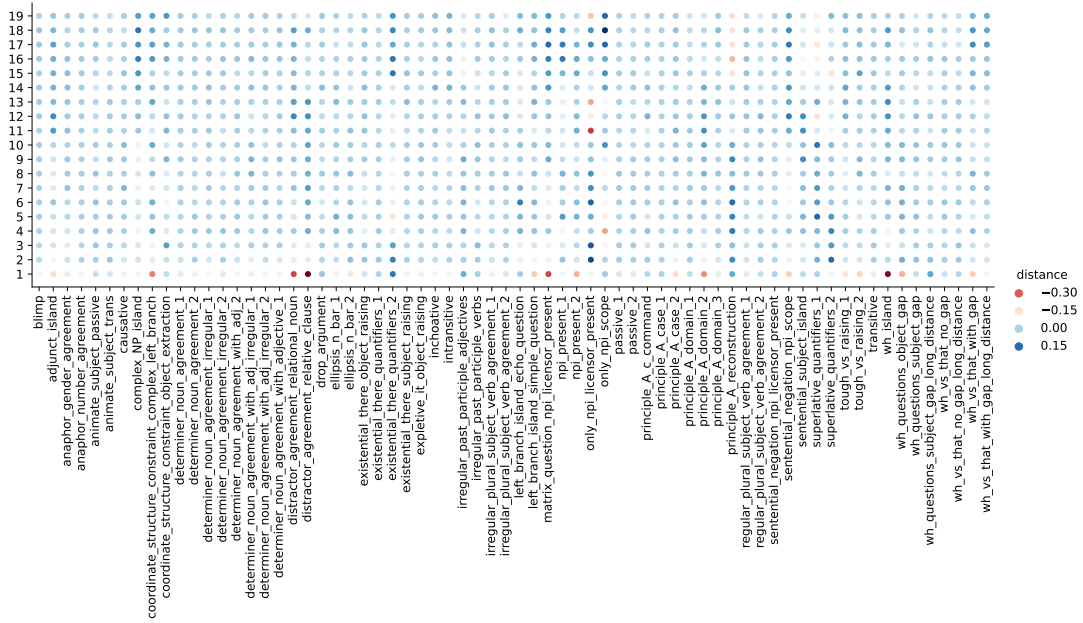


Figure 10: Paradigm-wise distances to mean paradigm performance for pythia-70m model

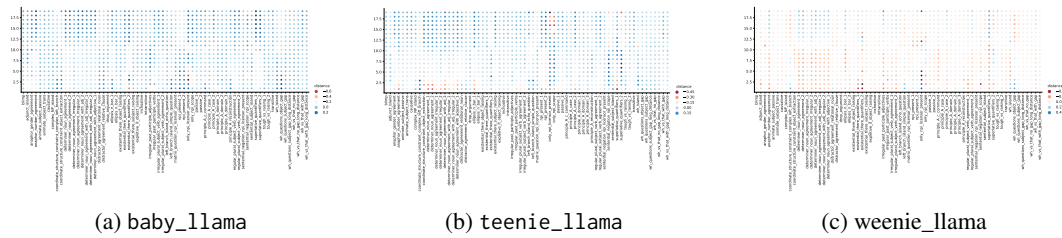


Figure 11: Distance to mean for baby_llama, teenie_llama and weenie_llama

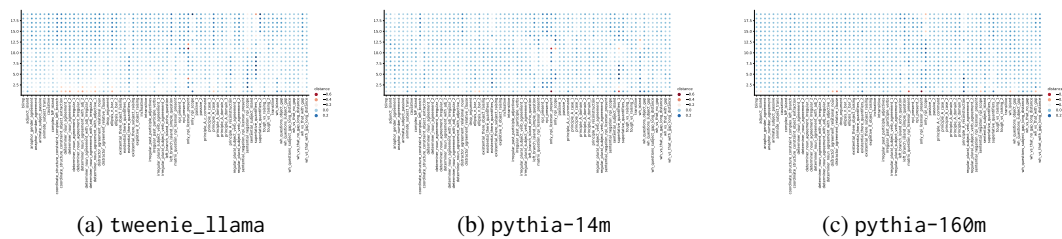


Figure 12: Distance to mean for tweenie_llama, pythia-14m and pythia-160m

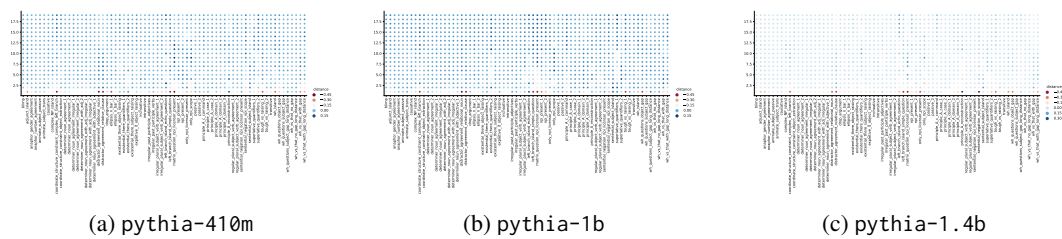


Figure 13: Distance to mean for pythia-410m, pythia-1b and pythia-1.4b