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

Bastian Bunzeck and **Sina Zarrieß**
Computational Linguistics, Bielefeld University, Germany
`{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de`

Abstract

Syntactic learning curves in LMs are usually reported as stable and power law-shaped. By analyzing the learning curves of different LMs on various syntactic phenomena using 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. Moreover, most phenomena show similar curves for their concrete paradigms, but the existence of diverging patterns and oscillations indicates that average curves mask important developmental differences.[1]

1 Learning curves

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". 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).

[1]Published at *Multimodality and Interaction in Language Learning (MILLing)*, see `https://aclanthology.org/venues/clasp/`

2 Methods

Models We analyze two different model architectures, four llama models (Touvron et al., 2023) trained on the 10M and 100M BabyLM 2023 data sets (Warstadt et al., 2023), and six pythia models (Biderman et al., 2023) trained on the much larger *The Pile* data set (Gao et al., 2020).

Evaluation We test linguistic knowledge as BLiMP performance with `lm-eval-harness` (Gao et al., 2022). BLiMP can be used to discern whether a grammatical sentence is preferred by an LM (lower perplexity): an accuracy of 50% equals the random baseline. We evaluate across the first training epoch and look at logarithmically spaced evaluation checkpoints: 10 checkpoints within the first 10% of training and 9 additional checkpoints until the epoch's completion.

Curves We devise our own classification scheme for learning curves based on the distinction between well- and ill-behaved curves in Viering and Loog (2023), and refine the categories with patterns attested in language acquisition (like U-shaped learning, see Saxton, 2009). Table 1 gives a brief overview. We qualitatively assign shapes to the learning curves, aided by fitting fifth-degree polynomials to each curve.

3 Results

The learning curves for all BLiMP phenomena and models are visualized in Figure 1. 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.

	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 1: Overview of proposed curve shapes

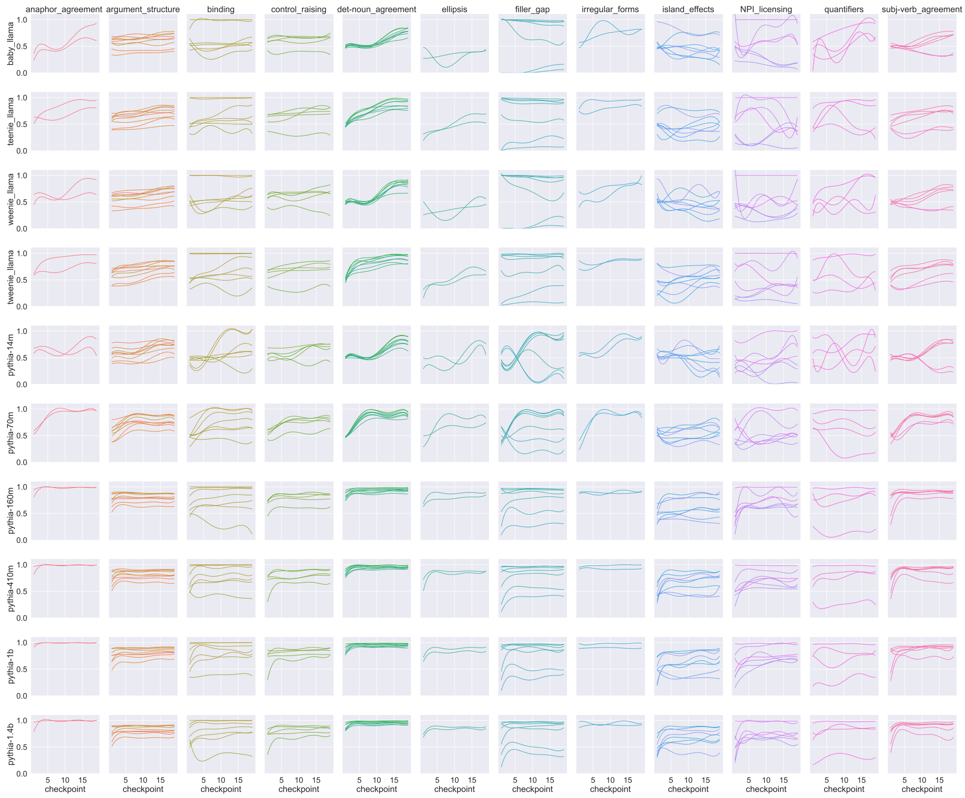

Figure 1: Learning curves for all paradigms in BLiMP, separated for models (rows) and phenomenon sets (columns)

- 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 these sheaves, also diverging patterns are observed within phenomena. Some paradigms within the same phenomenon have mirrored learning trajectories, where improvement in one paradigm correlates 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.

We conclude that while the rapid syntax learning assumption from earlier studies generally holds, it also needs revision. When averaging across many phenomena, performance gains seem to follow a prototypical power law. 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 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 is learned optimally.

References

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.

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.

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, Tttyuntian, 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/lm-evaluation-harness: V0.3.0. Zenodo.

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.

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.

Preetum Nakkiran. 2019. More Data Can Hurt for Linear Regression: Sample-wise Double Descent. *Preprint*, arxiv:1912.07242.

F.E. Ritter and L.J. Schooler. 2001. Learning Curve, The. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 8602–8605. Elsevier.

Naomi Saphra. 2021. Training dynamics of neural language models.

Matthew Saxton. 2009. The Inevitability of Child Directed Speech. In Susan Foster-Cohen, editor, *Language Acquisition*, pages 62–86. Palgrave Macmillan UK, London.

Peter Sollich. 2001. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

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. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

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.

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.