The geneticist Theodosius Dobzhansky famously declared, "Nothing in Biology Makes Sense Except in the Light of Evolution." In AI, we might say instead that "**Nothing in Deep Learning Makes Sense Except in the Light of Stochastic Gradient Descent**"—and my goal is to make sense of deep learning.

Our world increasingly relies on neural networks, but like biologists before Darwin, we are still at the stage of telling just-so stories to explain them. Supposedly interpretable artifacts are presented as explanations of model behavior, but we often are unable to determine whether a given model property is required for network performance or if it developed as a side effect of training. In biology, such non-functional properties might be *vestigial* traits or randomly selected *spandrels*, but **interpretability** and **science of deep learning** lack a framework for discussing similar phenomena.

I aim at a complete understanding of neural networks—primarily **language models** (LMs)—by characterizing how their internal mechanisms are constructed during **training**. This understanding can unlock many paths forward. We could debug models, verify their outputs, or predict possible operational failures. We might inspire improvements in architecture or optimization. Interpretation of generative models could also support new scientific understanding of the real world behind their training corpus. A *complete* understanding of model behavior would even enable simulation: Provided with training data and hyperparameter specifications, we could describe a hypothetical model's capabilities without having to train it.

My work, informed simultaneously by the sciences of linguistics and deep learning, has been at the forefront of a new wave of progress in interpretability. In LMs, interpretability research asks questions like: How do models internally represent language and world structure? What generalization behaviors and inductive biases are they likely to exhibit? What spurious shortcuts have they learned? Traditionally, these questions are asked of a fully trained model, an approach I have also contributed to with new probing methods [33, 17, 25]. However, by analyzing the training regime instead, my claims about structures and explanations hold beyond a single parameter setting; we can study the training *procedure*. How do models *learn* language and world structure? What generalization behavior does a *training method* exhibit? How are spurious shortcuts *discovered* and how can they be discouraged?

Past and current work

While many interpretability researchers ask *what* models learn, I ask *why* they learn. Every model behavior and mechanism has its origins in training; my objective is to explain how those training conditions lead to the resulting model. Given an architecture and optimizer, I view the training outcome as determined by three factors: **(1) the timescale of learning; (2) the composition of training data; and (3) random chance introduced by the initial seed**. Each of these factors has led me to new insights. By pursuing the complete picture, I have unlocked a new understanding of training that finally allows us to understand how and why LMs work.

Time

My approach centers the training process, particularly the timing of learning events. I developed this focus during my PhD, when I studied how the representations learned for next token prediction converged and diverged relative to representations targeting similar tasks like part-of-speech prediction [29]. I also attributed the powerful hierarchical biases of LSTMs to their bottom-up learning process [30]. These early papers proved prescient, inspiring others to study language model training.

Our more recent work on LM training has related abrupt phase transitions to specific model capabilities. We have revealed that a prolonged steep drop in the loss of masked LMs is actually composed of two consecutive breakthroughs [4]: The model first learns to internally represent linguistic structure with specialized attention heads, then learns to use that structure to follow complex grammatical rules (Figure 1). This finding was the first evidence of dependent conceptual breakthroughs in modeling real-world data, supporting my broad agenda of using training dynamics to aid model interpretation. Such breakthroughs may also be more common than previously documented; by decomposing changes in loss during training, we find [11] that individual gradient directions are associated with breakthroughs in specific skills, e.g., generating itemized lists, which are smoothed out by the aggregated training curve. This work has revealed **the interpretable nature and hidden ubiquity of these phase transitions.**

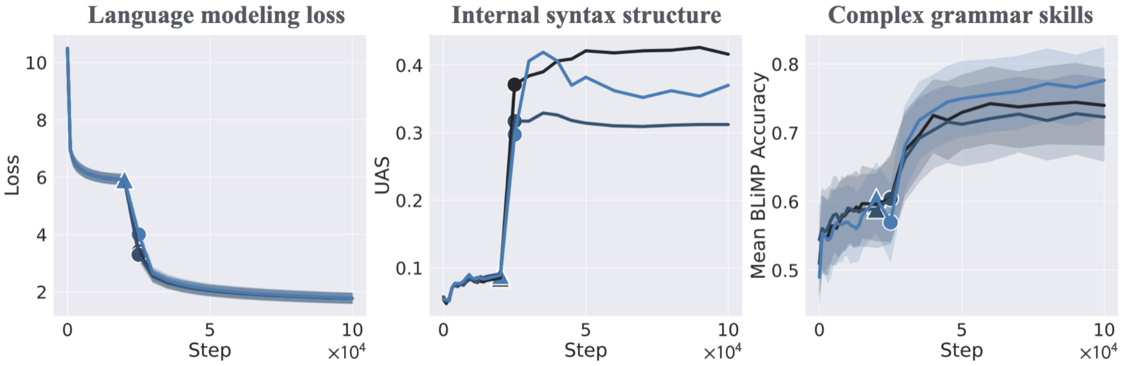

Figure 1: **Phase transitions in language model training.** In a masked LM, we discovered the multiple phase transitions underlying a prolonged steep drop in loss. Each color represents a different random seed. The **syntactic structure onset** (▲) precipitates the **grammar skills onset** (●) when the model suddenly learns to use its new internal structure to follow complex grammatical rules. Our findings confirm that specialized modules must form before the model even begins learning to use complex grammar, providing the first evidence of discrete dependencies in real-world artificial learning.

Data

What aspects of training data affect the final model? In addition to impacts from specific influential samples, e.g., sequence memorization (another topic I have studied [18]), overall data composition also affects whether a trained model can generalize in new **out-of-distribution** (OOD) settings. In multimodal visual question answering models, we showed that systematic compositionality, or the ability to combine atomic components in unseen ways, is induced by the diversity of contexts in which each atomic component was seen during training [2]. We even showed that diversity permits models to surpass the skills of the humans behind their training corpus, demonstrating theoretically and experimentally that diverse errors enable generative models to reflect the “wisdom of the crowds” in domains like chess [34].

In LMs, meanwhile, our work [19] uses English grammar acquisition to characterize how data complexity and diversity shape OOD behavior during training. As seen in Figure 2, LMs trained on syntactically homogenous examples memorize patterns, while one trained on diverse examples will learn general rules—but an intermediate level of diversity leads to unstable training, during which OOD behavior oscillates rather than committing to a rule. Data complexity shows similar patterns; complex sentences—examples with center embeddings—induce a correct hierarchical rule, while simple right-branching sentences induce n-gram-like linear rules. Mixing these subsets, again, destabilizes training. As data availability overtakes compute as the main resource bottleneck in training, **further performance gains rely on data-efficiency**. Whether that efficiency relies on data augmentation or curation, understanding the effect of corpus composition will only become more critical.

Chance

In machine learning, random variation is often disregarded, but we have found it to influence important LM behaviors like social bias [22]. Our previously mentioned work on data mixtures [19] revealed that, when the training data promotes an unstable regime, it also permits inconsistent generalization behavior across random seeds; stable training runs cluster around the two rule-based solutions. In work on finetuned text classifiers [10], we associated such clusters with different basins on the loss surface. These findings refuted the dominant belief that transfer learning leads inevitably to a single basin [5], a claim based solely on observations of image classifiers.

The clustering effects we observe in OOD generalization [10, 14, 19, 36] and training dynamics [32, 8] complicate narratives of sudden breakthrough capabilities at particular scales or training times; if each run must fail or succeed, with no continuous middle ground, then the associated capability can *only* emerge abruptly. In ongoing work [36], we reconcile the notion of continuous *scaling laws* in performance with sudden *emergent* breakthroughs

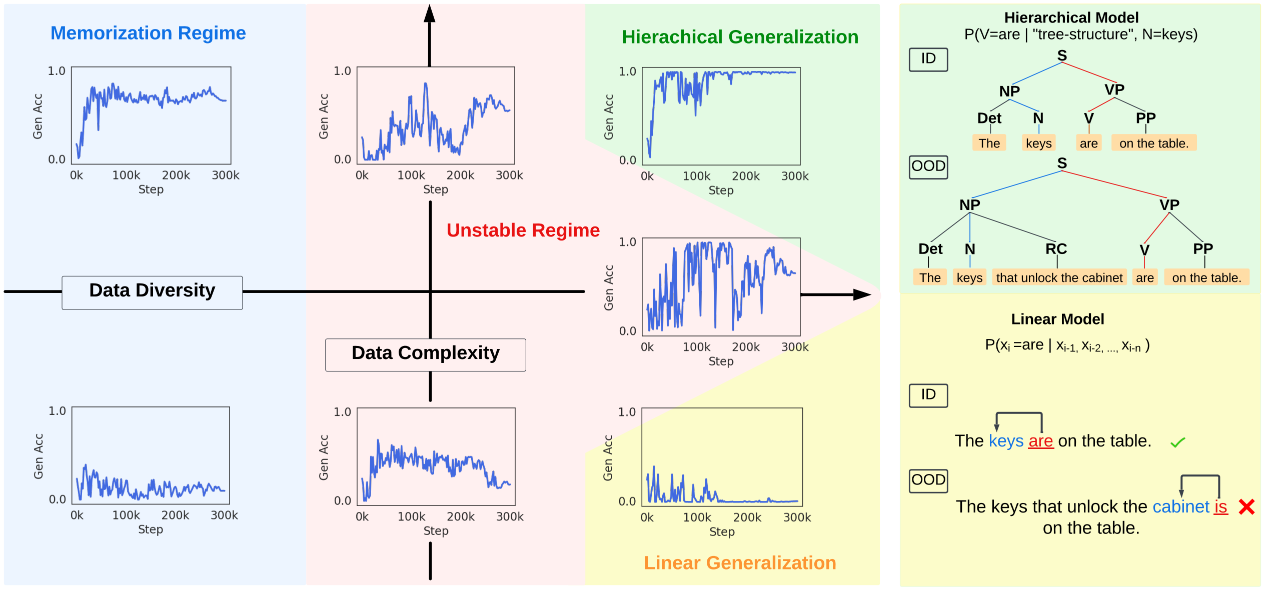

Figure 2: **Data determines OOD behavior.** Using the synthetic English setting of McCoy et al. [15], a model learns grammar rules from ambiguous data, compatible with both the correct hierarchical rule (*top*) and the n-gram-based linear rule (*bottom*). Diverse data teaches the model to generalize rather than memorize. Complex data induces the hierarchical rather than rule. Intermediate diversity or complexity leads to unstable OOD behavior.

on specific capabilities by studying continuous changes in the *probability* of a breakthrough across random seeds. By articulating random effects, we can **rigorously evaluate claims** about the impact of controlled variables like model size.

Next Steps

When I started asking these questions, there was little interest in understanding LM training. LM researchers ignored training dynamics while training dynamics researchers ignored LMs; many still-celebrated claims from the science-of-deep-learning literature apply only to image classifiers. But by advocating for LM analysis to include the training process [27], I changed the NLP community's culture.[1] My arguments were cited [22, 26] as justification when new LM releases first started including intermediate checkpoints for scientific study, now a standard open source practice [7, 3]. My work has drawn even more interest during the current interpretability research boom—in the last year alone, I have appeared on six workshop panels, often with hundreds in attendance.

Since I began publishing, there has been an explosion of papers on LM training. With the resulting new understanding of training, we can now improve LMs and predict their behavior in new environments.

Improving the training process through principled insights. Our understanding of the training process can guide new training methods for safer models, robustness, or greater efficiency. I have already used these insights to inform the design and evaluation of new methods for both pretraining [1] and finetuning [20, 23]. Next, I will consider human feedback post-training, where we are developing schedules for injecting fresh training data based on our recent counterintuitive findings about unstable phase transitions. Our existing findings also suggest that one can improve performance under data constraints by manipulating early latent structures. Our work on data

[1] My interest in cross-discipline parallels does not just inform my core research questions based on lessons from evolutionary biology. It has also inspired meta-scientific position papers aimed to stimulate debate and catalyze cultural change. Some document scientific history with implications for current research, e.g., parallels between the current LM scaling boom and the n-gram-based translation era [28]. Others highlight disciplinary divisions shaped by subculture [31] or geography [35]. Still others editorialize on evaluation [21] and epistemics [27]. These position papers are some of my most widely shared work; at least one has featured in a major conference keynote (EMNLP 2023).

construction and diversity can inform data augmentation research as well, by suggesting axes of diversity and complexity that most benefit the model when amplified. In the long term, I expect our scientific insights to yield further material gains in performance by developing new data practices, training schedules, and optimizers.

Unifying interpretability and evaluation. One rarely-attempted challenge in interpretability is predicting model behavior under specific distribution shifts. Although such an achievement would allow us to **use interpretation to assess model quality** by anticipating possible edge cases and biases, it is subject to numerous hurdles. Current interpretability methods cannot generalize OOD; even popular tools like sparse autoencoders can become useless in new contexts [16].

How could we then hope to solve this intractable challenge? We have found that random variation can form clusters in both model weights and OOD generalization behaviors [10], providing hope that by understanding the underlying mechanisms in-distribution, we can also predict outlier behavior. As a starting point, we have introduced a synthetic setting where random variation leads to different OOD behaviors as a testbed [14]. We can find precise mechanisms in-distribution that correlate with generalization rules, and even find specific markers that predict when these mechanisms will change their behavior on OOD data, potentially damaging the model's application of its rule [12]. When our synthetic work is complete, I will take on the challenge of predicting potential failure conditions in real-world models based on their internal mechanisms. One possible tool for this objective is information theory, which can anticipate reconstruction errors based on the model's compressed representations. With the resulting interpretability-based model evaluation, we could automatically identify potential deployment issues before we encounter them.

Long term: Interpretability for scientific understanding.

When the 2024 Nobel prize in chemistry was awarded to the protein folding model AlphaFold, it represented a broader recognition of AI as a method for scientific discovery. However, AlphaFold and other scientific discovery tools—now used to model phenomena in biochemistry, physics, or neuroscience—act like blackbox systems, forming predictions without providing any new human understanding. **The goal of scientific *discovery* is distinct from the goal of scientific *understanding*, but good enough *discovery* tools can empower *understanding* through model interpretation.** The interaction between LMs and the science of linguistics makes this promise clear.

Already, my work on LMs illuminates how language structure can be processed and learned, with consequences for linguistics. Our analysis of speech models [25] reveals that nonlinear interactions between acoustic input features align with how easily humans can understand a sound without its phonetic context. According to our work on data composition [19], exposure to syntactic center embeddings is necessary and sufficient for LMs to develop a preference for hierarchical grammatical rules, amplifying a key argument in the debate over the poverty of the stimulus, one of the oldest and most famous controversies in linguistics. LM analysis even provides insights into the underlying human population behind the training corpus; we have found that ChatGPT's guardrail system treats a simulated user more like a conservative if they endorse a football team with a more conservative fanbase [13].

My approach, however, can also reach scientific domains outside of linguistics and the social sciences by interpreting new scientific discovery models. These interpretations can be more principled if they are grounded in our understanding of model behavior as well as the target phenomenon; such an understanding must be informed by the training process. By referencing training, we can differentiate early simplistic heuristics from fully articulated rules and compare each to our existing scientific models, to better focus on our gaps in understanding.

I have already begun collaborating with scientists who want to understand their neural network models. In my collaboration with astrophysicists [6], we analyzed the internal representations of galaxy formation models to predict when models trained under simulation conditions would fail to generalize OOD, with the ultimate goal of improving generalization to real-world data. In my ongoing collaboration with neuroscientists and neuroethologists [9, 24], we study models of weakly electric fish behavior to understand animal communication and group foraging. In the long term, these collaborations are where I see my work having the greatest impact—shedding light on natural phenomena we do not yet understand. Given that my thinking has been shaped by the work of evolutionary biologists, I look forward to shaping the progression of the natural sciences as well.

References

[1] Zachary Ankner, **Naomi Saphra**, Davis Blalock, Jonathan Frankle, and Matthew L. Leavitt. Dynamic masking rate schedules for MLM pretraining. In *European Association for Computational Linguistics (EACL)*, 2024. URL https://arxiv.org/abs/2305.15096. Accepted as oral presentation.

[2] Ian Berlot-Attwell, Kumar Krishna Agrawal, A. Michael Carrell, Yash Sharma, and **Naomi Saphra**. Attribute diversity determines the systematicity gap in VQA. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2311.08695.

[3] 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. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

[4] Angelica Chen, Ravid Schwartz-Ziv, Kyunghyun Cho, Matthew Leavitt, and **Naomi Saphra**. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=MO5PiKHELW. Spotlighted (top 5%).

[5] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis, 2020. URL https://arxiv.org/abs/1912.05671.

[6] Yash Gondhalekar, Sultan Hassan, **Naomi Saphra**, and Sambatra Andrianomena. Towards out-of-distribution generalization in large-scale astronomical surveys: robust networks learn similar representations. In *NeurIPS workshop on Machine Learning and the Physical Sciences*, 2023. URL https://arxiv.org/abs/2311.18007.

[7] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hanna Hajishirzi. Olmo: Accelerating the science of language models. *arXiv preprint*, 2024. URL https://api.semanticscholar.org/CorpusID:267365485.

[8] Michael Hu, Angelica Chen, **Naomi Saphra**, and Kyunghyun Cho. Delays, detours, and forks in the road: Latent state models of training dynamics. *Transactions of Machine Learning Research (TMLR)*, 2023. URL https://arxiv.org/abs/2308.09543.

[9] Sonja Johnson-Yu, Satpreet Harcharan Singh, Federico Pedraja, Denis Turcu, Pratyusha Sharma, **Naomi Saphra**, Nathaniel Sawtell, and Kanaka Rajan. Understanding biological active sensing behaviors by interpreting learned artificial agent policies. In *Workshop on Interpretable Policies in Reinforcement Learning @RLC-2024*, 2024. URL https://openreview.net/forum?id=FX7YtfEYj8. Accepted as oral presentation.

[10] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and **Naomi Saphra**. Linear Connectivity Reveals Generalization Strategies. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2205.12411.

[11] Sara Kangaslahti, Elan Rosenfeld, and **Naomi Saphra**. Loss in the crowd: Hidden breakthroughs in language model training. In *ICML Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=Os3z6Oczuu. Spotlighted. Under submission to ICLR 2025.

[12] Jenny Kaufmann*, Victoria R. Li*, Martin Wattenberg, David Alvarez-Melis, and **Naomi Saphra**. Causation does not imply correlation: A study of circuit mechanisms and model behaviors. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

[13] Victoria R. Li*, Yida Chen*, and **Naomi Saphra**. ChatGPT doesn't trust Chargers fans: Guardrail sensitivity in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL `https://arxiv.org/abs/2407.06866`.

[14] Victoria R. Li*, Jenny Kaufmann*, David Alvarez-Melis, and **Naomi Saphra**. Twin studies of factors in OOD generalization. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

[15] R Thomas McCoy, Robert Frank, and Tal Linzen. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 2020.

[16] Abhinav Menon, Manish Shrivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing (in)abilities of SAEs via formal languages, 2024.

[17] Tiago Pimentel*, **Naomi Saphra***, Adina Williams, and Ryan Cotterell. Pareto Probing: Trading Off Accuracy for Complexity. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL `https://aclanthology.org/2020.emnlp-main.254/`.

[18] USVSN Sai Prashanth*, Alvin Deng*, Kyle O'Brien*, Jyothir S V*, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and **Naomi Saphra**. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon, 2024. URL `https://arxiv.org/abs/2406.17746`. Under submission to ICLR 2025.

[19] Tian Qin, **Naomi Saphra**, and David Alvarez-Melis. Sometimes I am a tree: Data drives fragile hierarchical generalization. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning* and *NeurIPS Workshop on Compositional Learning*, 2024. URL `https://openreview.net/forum?id=juxbsQEuTZ`. Under submission to ICLR 2025.

[20] Adir Rahamim, **Naomi Saphra**, Sara Kangaslahti, and Yonatan Belinkov. Fast forwarding low-rank training. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL `https://arxiv.org/abs/2409.04206`.

[21] Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and **Naomi Saphra**. Benchmarks as microscopes: A call for model metrology. In *Conference on Language Modeling (COLM)*, 2024. URL `https://arxiv.org/abs/2407.16711`.

[22] Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, **Naomi Saphra**, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://arxiv.org/abs/2106.16163`. Spotlighted (top 5%).

[23] Tom Sherborne, **Naomi Saphra**, Pradeep Dasigi, and Hao Peng. TRAM: Bridging Trust Regions and Sharpness Aware Minimization. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2310.03646`. Spotlighted (top 5%).

[24] Satpreet Harcharan Singh, Zhouyang Lu, Kanaka Rajan, Sonja Johnson-Yu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, **Naomi Saphra**, and Nathaniel Sawtell. Investigating active electrosensing and communication in artificial fish collectives, 2024. Under submission to Cosyne 2025.

[25] Divyansh Singhvi*, Andrej Erkelens*, Raghav Jain*, Diganta Misra, and **Naomi Saphra**. Shapley interactions for complex feature attribution. In *NeurIPS Workshop on Attribution at Scale*, 2023. URL `https://openreview.net/forum?id=R3Lcbm7BcX`. Under submission to ICLR 2025.

[26] Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.11.

[27] **Naomi Saphra**. Interpretability creationism. *The Gradient*, 2023. URL `https://thegradient.pub/interpretability-creationism`.

[28] **Naomi Saphra**, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. First tragedy, then parse: History repeats itself in the new era of large language models. In *North American Association for Computational Linguistics (NAACL)*, 2024. URL `https://arxiv.org/abs/2311.05020`.

[29] **Naomi Saphra** and Adam Lopez. Understanding Learning Dynamics Of Language Models with SVCCA. In *North American Association for Computational Linguistics (NAACL)*, 2019. URL `https://aclanthology.org/N19-1329/`.

[30] **Naomi Saphra** and Adam Lopez. LSTMs Compose—and Learn—Bottom-Up. In *Findings of Empirical Methods in Natural Language Processing (EMNLP Findings)*, 2020. URL `https://aclanthology.org/2020.findings-emnlp.252/`.

[31] **Naomi Saphra*** and Sarah Wiegreffe*. Mechanistic? In *EMNLP BlackboxNLP Workshop*, 2024. URL `https://arxiv.org/abs/2410.09087`. Accepted as oral presentation.

[32] Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, **Naomi Saphra**, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs, 2024. URL `https://openreview.net/forum?id=bmrYu2Ekdz`. Under submission to ICLR 2025.

[33] Jennifer C. White, Tiago Pimentel, **Naomi Saphra**, Adina Williams, and Ryan Cotterell. A Non-Linear Structural Probe. In *North American Association for Computational Linguistics (NAACL)*, 2021. URL `https://arxiv.org/abs/2105.10185`.

[34] Edwin Zhang, Vincent Zhu, **Naomi Saphra**, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham M. Kakade, and Eran Malach. Transcendence: Generative models can outperform the experts that train them. In *Neural Information Processing Systems (NeurIPS)*, 2024. URL `https://arxiv.org/abs/2406.11741`.

[35] Bingchen Zhao*, Yuling Gu*, Jessica Zosa Forde, and **Naomi Saphra**. One Venue, Two Conferences: The Separation of Chinese and American Citation Networks. In *NeurIPS Workshop on Cultures of AI and AI for Culture*, October 2022. URL `https://openreview.net/forum?id=9U6w1tSZu_T`.

[36] Rosie Zhao, Sham Kakade, and **Naomi Saphra**. Distributional scaling laws for emergent capabilities. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning*, 2024.