While most of my teaching experience comes in the form of small interactive tutorial sections for advanced courses, I have also mentored individual students in their research and performed large-scale outreach to audiences outside of machine learning—even outside of computer science entirely. At every scale, I remove access barriers by making the topic entertaining and engaging for those new to the field.

Deep learning practice, NLP, and machine learning. An introductory course in language models (LMs) should teach three things: what current models are capable of, how they work, and why they work. The first teaches the student to be an informed user; the second, an engineer; the third, a scientist.

To be an informed user, the student must have direct experience with the latest models and an understanding of their power and limitations. A flood of AI hype has triggered a backlash of reflexive dismissal, but an informed user questions both default positions and knows where LMs still underperform. The value of this knowledge is twofold: to inform students as consumers and to motivate improvements.

To be an engineer, students must execute an LM-based project. Libraries like pytorch simplify implementation, but execution still introduces challenges when training models on real world data. Debugging requires an understanding of current architectures and optimizers. Expertise requires intuition around machine learning theory and empirics, e.g., ideas about distribution shift, overfitting, and double descent.

The last challenge is to make students scientists. Clearly, the current AI boom supports the notion that simple methods at scale outperform data-informed methods. However, it also subtly subverts sweeping claims about the obsolescence of domain knowledge: Although architecture and training methods may be data-independent, trained models must leverage language structure. Students should therefore understand the properties of language that make LMs powerful—compositionality and information economy.

I hope to teach a curriculum that unifies deep learning practice, NLP, and machine learning. I am, however, equipped to teach each individual topic separately; most of my lecturing experience is in math-heavy machine learning courses like *Probabilistic Modeling and Reasoning* and *Machine Learning and Pattern Recognition*. As well as introductory NLP/LM courses, I can teach advanced seminars on AI **interpretability**, **LM training dynamics**, or **science of deep learning**. Regardless of the course objective, I will discuss recent breakthroughs to ground student interest.

Engagement. Instruction should be engaging, especially for new students of machine learning and for established scholars in other disciplines. While it can be easy to share excitement about specific technical results with advanced practitioners, I also want to make AI accessible to scholars in art, science, and the humanities. To that end, I have targeted communities that connect with my research on a purely political or philosophical level, including presenting a paper on automation at a summit about society's changing relationship with work [1]. Having explained federated machine learning algorithms to this group of economists, journalists, and political theorists, I could easily keep the attention of any audience already interested in machine learning.

Whether I present to machine learning novices or outsiders, playfulness and humor are critical tools. As an undergraduate, I helped build a curriculum for local middle school girls, using play to introduce topics like overfitting or regular expressions. At a comedy club in Edinburgh, I performed a standup set about machine learning that has since been watched thousands of times and excerpted on Gary Marcus's Humans vs Machines podcast. My outreach skills are transferable to teaching generally, as lecturing at any level is fundamentally a performance requiring audience engagement. **As part of my commitment to engaging lectures, I attend courses on clown performance taught by professional clowns.**

Individual mentorship. I often rely on a core insight from clown pedagogy: that moments of friction are opportunities. The best setting to take advantage of friction, in the form of student confusion or

questions, is through individual mentorship where I can make students feel at ease to express uncertainty.

In addition to several early PhD students whose work I have supervised, I have advised three master's students in their thesis work [2, 3, 4] and six in their capstone projects. I am especially proud of my undergraduate mentorship, including several published papers with first authors who began as undergraduates [5, 6, 7]. I initially publicize these projects, but ensure student ownership in the final official presentation. This combination of security and independence has served students well. One undergraduate coauthor even asked me whether I would be starting as faculty soon so they could apply to my lab. Another longtime collaborator told me that my presence was a factor in him accepting a PhD offer at Harvard. I aim to enable students' success while imbuing them with a passion for research, and thus far the results have been exceptional.

References

[1] Kate McCurdy and **Naomi Saphra**. Carbon AI and the concentration of computational work. In *Challenging the Work Society: an interdisciplinary summit*, 2019.

[2] Alp Ozkan. Combined application of pruning and growing approaches for neural networks. Master's thesis, University of Edinburgh, 2018. Co-advised with Adam Lopez.

[3] Sylke Gosen. Understanding language models through perturbed datasets. Master's thesis, University of Amsterdam, 2021. Co-advised with Dieuwke Hupkes and Jaap Jumelet.

[4] Yekun Chai. Discovering spelling variants on urban dictionary. Master's thesis, University of Edinburgh, 2018. Co-advised with Adam Lopez.

[5] 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. Accepted as oral presentation.

[6] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and **Naomi Saphra**. Linear Connectivity Reveals Generalization Strategies. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] 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.