I'm interested in the **basic science of foundation models** that understand and interact through language, speech and vision; the **evaluation methods** necessary to answer scientific questions about these models; and how **practical applications** inform basic science about foundation models, particularly in technical, scientific fields. Below, I outline interests I would like to pursue in my graduate work:

Experimental practice for language modeling. A material scientist can discover a new alloy without building a bridge, but they *need to know how a bridge is built*. Similarly, a graduate student can make discoveries about language modeling without training an language model, but they *need to know how language models are trained*. I'm interested in open recipes for foundation models, with a focus on methodology as a distinct object of study: studying how scientists make empirical claims about the models they build, and how their hypotheses interact with evaluation, scaling and data.

For instance, a researcher pretraining a language model relies on small-scale training runs to test changes to data, architecture or training procedures, and ultimately selects the top-ranking configurations for their large-scale run. Yet in practice, experimental noise is often far too large to detect whether small-scale findings will generalize [1]. In my **NeurIPS'25 Spotlight** paper, I found a simple indicator of the likelihood that an experimental result would hold from small to large compute scales: the ratio of *signal*, a benchmark's ability to separate better models from worse models, and *noise*, a benchmark's sensitivity to random variability between training steps [2]. Naturally, my work enabled more precise decision-making about pretraining LMs: with *task scaling laws* (**COLM'25**) we proposed that the downstream performance of large models was predictable, for some tasks within 2% relative error (e.g. MMLU), by fitting a task scaling law to a set of small 'ladder' models trained with less than 1% of the compute of the large run [3]; and my later work on *DataDecide* (**ICML'25**) studied model selection of LMs trained using different datasets, showing the compute scale necessary to distinguish between data methods and suprisingly performant proxy metrics for making decisions [4]. By focusing on better methods, these efforts have shown that language modeling does not need to be an alchemy: each work articulated one aspect of evaluation, scaling and data. In my graduate work, I'm interested in how scientists design claims beyond language modeling alone: models that learn from images, videos, human speech and interaction. Take late-fusion vision- or speech-language models – we have made progress on new architecture ideas, yet standard data and evaluation practices have not been established, which has made it unclear whether basic decisions actually work (e.g. how to align text and speech position embeddings or how exactly to patch images and video).

Methodology work should also *actually improve* how we make scientific claims about our models. As a core contributor to **Olmo 3**, I led the experimental design used by the pretraining team in 1000s of preliminary runs: task scaling laws and the DataDecide suite allowed us to select the ablation configuration for validating data choices (like quality upsampling) and modeling choices (like SWA) before the large run [5]. When I designed the evaluation suite, I optimized for a high *signal-to-noise ratio*, making it possible to fit the Olmo 3 data mix with models as small as 30M and even on 'emergent' math and code abilities [6]. During my Ph.D., I aim to use method work to answer new hypotheses across the modeling pipeline. E.g., what if our scaling laws could detect *early failure*, by predicting validation loss for an entire training curve? Could we assess data's value to model capabilities without training any models? As a timely example, given the recent attention to scaling RL training, little has been said about data: Can we characterize environments as data and apply data-centric pretraining methods (mixing, task scaling laws) to scale RL? I'm eager to develop scientific principles for foundation models to answer new questions like these.

The *how* of evaluating language models. The existing crisis of *what* to evaluate – defining the capabilities we want through new and harder benchmarks – has an adjacent problem: *how* we evaluate, which I expect to grow in importance as our models become more capable. We now evaluate tasks which execute for hours and use multi-million token contexts [7], yet only provide a binary judgment of correctness. At the same time, the difficulty in collecting expert-labeled benchmarks have resulted in a reliance on less statistical power: AIME 2025, used in over 50 NeurIPS papers this year, consists of 30 questions. I think new ideas are needed in evaluation design: I'm excited about methods to provide a precise definition of error, an understanding of why failure occurs, and a clear direction to improving models. As a first step, I helped introduce *Fluid*

Benchmarking (**COLM'25, Oral**), a method based on item response theory, which selects evaluation items dynamically to maximize the information gain of each instance [8]. My experiments show that for many benchmarks, e.g. GSM8K, the learned item response model exhibited a lower variance of scores despite using 50x fewer items; and for others, e.g. MMLU, the method automatically avoided mislabeled questions. By revisiting basic assumptions about model benchmarks, I aim for measures that dynamically match capability of the model being tested, prioritize worst-case reliability and allow expressing uncertainty. For example, a scientist may prefer a model with a high abstention rate in exchange for statistical guarantees that the model does not hallucinate; our evaluation should account for this.

I'm also interested in evaluation with greater explanatory power, particularly formulations more expressive than binary outcomes. In work I led on *edit-level evaluation* (**EMNLP'23**), we introduced a new evaluation approach by annotating text spans corresponding to desirable or undesirable behavior in language model generation [9]. We worked with human annotators to define a typology of 21 edit types in text simplification, which we used to train a set of automatic evaluation metrics using a novel dual token- and sentence-level objective. This evaluation provides a single score *and* a set of error types corresponding to specific text spans in the model's generation, a level of granularity previously only captured qualitatively. Our follow-up demo paper (**EMNLP'23 Demo**) released an easy-to-use package for span-based language model evaluation [10], which was used in human studies of medical factuality and idiom translation, among others. At the TSAR workshop panel at EMNLP'24, I discussed how this span-based evaluation regime also allows individuals to express errors specific to their desired simplification level – I believe measures of error which account for the needs of *individual users* is the most important missing quality of our benchmarks.

Assisting scientific discovery with language models. There is a mismatch between the tasks receiving focus from researchers (e.g. competition math and coding) and capabilities we actually seek in a useful AI system. Looking forward, I'm interested in applications that provide direct, practical value to the general public. I find assisting scientific discovery a particularly exciting direction: I think a system useful for scientific work will have the form factor of a language model agent which can operate on the same data and tooling that scientists use. However, while contributing to *Terminal-Bench* [7], I realized agents are still very far from something that scientists would find useful in their daily work. In an ongoing project at Ai2, I've been training LM agents to assist with one aspect of scientific work – reproducibility – by designing new training objectives for replicating empirical claims in computational papers. We've found new problems: in one experiment using a sample of ACL papers, roughly 2/3 no longer have online the published code, data and models necessary for reproducing their findings. This work has led me to believe that scientific discovery asks new questions about the entire modeling stack: the training objective we eventually design will need to handle experimental uncertainty in the same way scientists do. The evaluation regime will need to measure how our models impart *understanding* in the scientists that use them, rather than measuring how well we can automate experimentation. Starting with a better understanding of this family of capabilities, I aim to inform new, use-inspired methods which change how foundation models are developed.

References

[1] Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, ..., **David Heineman**, ..., Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious. In *COLM*, 2025.

[2] **David Heineman**, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *NeurIPS*, 2025.

[3] Akshita Bhagia*, Jiacheng Liu*, Alexander Wettig, **David Heineman**, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *COLM*, 2025.

[4] Ian Magnusson*, Nguyen Tai*, Ben Bogin*, **David Heineman**, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *ICML*, 2025.

[5] Olmo Team (incl. **David Heineman**). Olmo 3. *preprint*, 2025.

[6] Mayee Chen, Tyler Murray, **David Heineman**, Matt Jordan, Hannaneh Hajishirzi, Christopher Ré, Luca Soldaini, and Kyle Lo. Olmix: Efficient mixture recomputation for evolving LM datasets. *preprint*, 2025.

[7] Mike A Merrill, Alexander Glenn Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, ..., **David Heineman**, ..., Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *preprint*, 2025.

[8] Valentin Hofmann, **David Heineman**, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. In *COLM*, 2025.

[9] **David Heineman**, Yao Dou, and Wei Xu. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *EMNLP*, 2023.

[10] **David Heineman**, Yao Dou, and Wei Xu. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In *EMNLP: System Demonstrations*, 2023.