

Research Statement

Nihal V. Nayak

I study **zero-shot generalization in machine learning, the ability of a system to solve new tasks without human annotations for those tasks [10]**. Frontier AI systems, such as Claude and ChatGPT, are built by training large machine learning models on enormous amounts of unstructured internet-scale data along with data annotated by human experts. This training strategy has demonstrated remarkable *zero-shot* capabilities to solve new tasks in complex environments. But these systems often fail unpredictably, and detecting their failures at scale is quickly becoming a challenging task [24]. Furthermore, systematically addressing these failures, rather than relying on ad hoc fixes [18], often requires costly retraining with substantial compute and capital, highlighting the need for efficient training methods. **My work introduces scalable, data-centric learning techniques to efficiently supervise and audit frontier AI systems for better zero-shot generalization [2–5, 7, 9, 10, 12, 13, 16, 17]**. My work lays important groundwork for building the next generation of reliable AI systems. Below, I have summarized my research:

- **Principled Data Curation:** I have worked on principled data curation since the early days of LLM post-training, from handcrafting instruction-tuning datasets to generating synthetic datasets at scale [2, 12, 16]. Building on this experience, my recent work addresses two central challenges: selecting useful training examples and generating high-quality supervision for specialized domains. I introduced a framework that disentangles the key components of data selection in LLM post-training [13]. Through systematic experiments and generalization bounds, I showed **how to select data that predictably and provably improves downstream performance** and when inexpensive random sampling is competitive. For synthetic data generation, **I created Bonito, the first open-weight LLM for conditional task generation**, which converts users’ corpora from specialized domains into synthetic instruction-tuning datasets [12]. Bonito enables practitioners to rapidly generate instructions to train specialized language models without human annotations.
- **Efficient Pre-Training:** Deploying LLMs across diverse hardware requires models of different sizes, but training each model independently is prohibitively expensive. In collaboration, **I discovered boomerang distillation, a simple technique that creates LLM families of fine-grained sizes through model size interpolation [5]**. By progressively reintroducing teacher layers into a distilled student, we created intermediate-size models whose performance smoothly interpolates between the student and teacher without additional training while requiring **14-19× fewer training FLOPs** than individual distillation.
- **Auditing Generalization:** Auditing the generalization of AI systems along with the training recipes is becoming increasingly challenging. I use techniques and concepts from Cognitive Science to create diagnostics to evaluate AI systems [7, 9]. To systematically study the limits of compositional generalization in vision-language models, I created the concept binding benchmark [9], a synthetic dataset that tests if vision-language models can identify different objects and reason about their properties individually. **I showed that foundation models such as CLIP critically fail when classifying compositions with multiple objects.**

Impact. My work has been published in top machine learning and natural language processing (NLP) conferences and journals [1–6, 9, 11–16], including a Spotlight at ICLR. My work on synthetic data generation [12] was instrumental in **securing an NSF grant [23] worth over \$660K to build AI systems** for scientists. My collaborative work on training large language models with prompts to follow instructions — T0 [16] and BLOOM [3] — laid the foundations for today’s frontier AI systems. I have also **open-sourced several frameworks and models that practitioners have downloaded thousands of times** to improve their AI systems. My work has gained recognition across academia and industry, with coverage in Nature [21], Thoughtworks Report [25], and well-known technology blogs [19, 22].

Why Academia? We are at a pivotal point, witnessing the real-world impact of decades of machine learning research and realizing AI’s true promise: to benefit humanity by accelerating innovation and enabling new discoveries. I believe academia is uniquely positioned to address long-term research goals without the financial pressures of shareholders. We have the exciting opportunity to contribute to the **open science of building reliable frontier models** and share our findings openly with the community. In the future, **I will continue to develop cost-effective, scalable, data-centric solutions to build the next generation of reliable AI systems.**

Principled Data Curation

Efficiently building the next generation of AI systems requires automated methods to select and generate high-quality training data that improve their capabilities. My recent work advances this goal by identifying how to select training data that predictably improves downstream performance on target tasks and how to synthesize training datasets from unstructured expert documents.

Predictably Selecting High-Quality Instruction Tuning Datasets

Instruction tuning is a critical step in the LLM post-training pipeline, where the pretrained LLM is trained to follow instructions [3, 7, 16]. However, post-training LLMs often require elaborate data-curation pipelines that repeatedly train LLMs by ablating over datasets with different ratios of task types, styles, and domains. A key question is **how can we efficiently select high-quality training datasets to post-train LLMs?** Existing frameworks offered limited practical guidance because they differed in their selection mechanisms and training budgets, often conflated the contributions of individual components, and did not reliably achieve strong downstream performance.

I developed a framework that disentangles the key components of automated data selection methods — data representations and selection algorithms — to isolate their contributions and provide efficient training recipes for practitioners [13]. I introduced a diagnostic to test which data representations make selecting training subsets closer to the target task a reliable predictor of lower prediction error. I found that **only gradient representations achieve near-perfect Spearman correlation between subset distance and target-task loss**: training on subsets ranked as closer to the target task consistently yields lower loss. Text-based representations often show little to no correlation, making their performance gains less predictable. Then, I showed that combining these gradient representations with greedy round-robin selection often yields the best performance at small budgets, measured by the number of selected training examples. For larger budgets, I developed an **unbalanced optimal transport selection algorithm** that offers better performance. Surprisingly, random selection often matches or outperforms popular selection methods as training budgets increase, helping practitioners assess when targeted selection offers gains that justify its computational cost. Finally, I unified the selection algorithms studied as forms of approximate distance minimization between the selected subset and the target task, and **derived generalization bounds, based on optimal transport, that connect target-task loss to subset-target task distance and characterize the budget-dependent limits on gains over random sampling**. These bounds help explain why the advantages of targeted selection can diminish as training budgets grow. This work provides a natural framework for evaluating new selection methods, identifying which design choices drive improvements, and guiding the development of more reliable and efficient approaches to post-training LLMs.

Learning to Generate Synthetic Instruction Tuning Datasets

Distillation is a commonly used technique to create synthetic instruction tuning datasets by querying the API end point of a frontier model and then training a smaller model. Often organizations want to adapt general-purpose LLMs to perform well on their custom tasks in their specialized domain but cannot create synthetic instruction tuning data through distillation as it requires sharing private and proprietary data with the API-based model. This raises an important question: How can we adapt general-purpose language models to specialized domains without distilling from an API-based model? To address this problem, I created **Bonito, the first open-weight language model that converts a user’s unannotated domain-specific corpus into synthetic instruction-tuning datasets without human annotations** [12]. Bonito enables organizations to generate these datasets within their own infrastructure, supporting domain adaptation without sending sensitive corpora to external services. I trained Bonito by fine-tuning a pretrained language model on 1.65M examples created by “remixing” existing instruction-tuning datasets, such as P3 [2], into *meta-templates*. These meta-templates produce training examples that map unannotated text and a task attribute to an instruction and its response. My experiments showed that Bonito-generated datasets significantly improve the performance of several pretrained and instruction-tuned language models over standard baselines across specialized domains. The project has **received over 800 stars on GitHub¹, and Bonito models have been downloaded over 20,000 times** on Hugging Face. Bonito provides practitioners and domain experts with an accessible path to adapting open-weight models to their own data without manual annotation. My work has gained recognition across academia and industry, with coverage in the Thoughtworks Report [25] and well-known technology blogs [19, 22].

¹<https://github.com/BatsResearch/bonito>

Efficient Pre-Training

Pretrained models, particularly large language models, are the backbone of every intelligent system, but they are often rigid and come in coarse-grained size families, which prevents rapid deployment in heterogeneous settings. In this work, we show how to interpolate between model sizes to create a family of large language models of fine-grained sizes [4, 5, 17].

Boomerang Distillation Enables Zero-Shot Model-Size Interpolation

Adapting large language models to diverse hardware and deployment constraints remains a critical challenge as these models are deployed in settings ranging from edge devices to large-scale clusters [6]. To meet these diverse requirements, model providers often release models at several scales, collectively referred to as a model family. However, creating LLM families remains highly resource-intensive because conventional pipelines often require a separate training or distillation run for each model. In collaboration, I discovered **boomerang distillation** [5], a phenomenon that enables zero-shot model-size interpolation and **creates LLM families of fine-grained sizes with just a single training run**. We showed that distilling a large base model (the teacher) into a smaller student model and then progressively reintroducing teacher layers to patch the student produces intermediate-size models without additional training. Across classification and generation tasks, we showed that the performance of these zero-shot interpolated models varies smoothly between that of the student and teacher models. Compared with distilling each model individually, **boomerang distillation requires 14-19× fewer training FLOPs** while matching the performance of individually distilled models at smaller sizes and outperforming them at larger sizes.

To further improve the performance of the interpolated models, we investigated the role of layer-patching order [4], that is, which student layers should be replaced with teacher layers. We proved that finding the optimal patching order can be formulated as a shortest-path problem on a directed acyclic graph and proposed **KLPatch, a greedy approximation based on KL divergence that reduces the computational complexity from $O(N!)$ to $O(N^2)$** and improves downstream performance. Finally, we extended boomerang distillation to post-trained models [17], demonstrating that the cost of **a single distillation run can be amortized across K post-trained variants**, including instruction-tuned and reasoning models, without additional training or adaptation. Together, these contributions lay the foundation for a new generation of flexible models that can adapt to real-world hardware and resource constraints.

Auditing Generalization

Auditing the generalization of AI systems is becoming increasingly challenging. Therefore, we need tools and evaluation methods that help us identify the critical limitations of the technology. I use techniques and concepts from Cognitive Science to develop diagnostics that identify critical limitations of our existing training recipes [7] and foundation models [9].

Concept Binding in Vision-Language Models

Compositionality is a long-standing goal of artificial intelligence: creating new concepts by combining existing primitive concepts. Evaluating compositionality in vision-language models such as CLIP is challenging because these models are trained on opaque, internet-scale data, and a positive result could stem from a leak in the training data or a spurious correlation. Furthermore, existing evaluation datasets that test for compositionality are often hackable [27], as one can achieve very high scores on their vision-language benchmark even without looking at the image. To address these issues, I created **the concept binding benchmark**, a synthetic dataset to evaluate compositionality in vision-language models [9]. This benchmark draws on a cognitive science phenomenon called *binding* [26]: the fundamental ability of humans to associate each attribute with the correct object. For example, an image in the benchmark contains a red cube and a yellow sphere; correctly identifying “red cube” requires binding each color to the appropriate shape and rejecting plausible distractors such as “red sphere” and “yellow cube”. Unlike prior work, which is susceptible to spurious correlations and hackable artifacts, the synthetic benchmark provides a controlled setup to test compositionality in three key settings—single-object, two-object, and relational—while ensuring the dataset provides semantically meaningful and hard distractors. Upon evaluation, I found that CLIP and its fine-tuned variants, including my prior work that efficiently improves compositionality in VLMs [14], perform well on single-object adjective and noun compositions, but **all models show near-chance performance on two-object and relational compositions that require concept binding**.

These findings reveal a critical limitation in CLIP’s representation of multi-object images that may affect downstream technologies built on top of it. Later studies highlighted this issue across other types of models such as diffusion models [20].

Future Work

In the future, I will continue to develop data-centric solutions for building the next generation of reliable frontier AI systems. I strongly believe that the **next generation of AI systems will self-improve and self-adapt** to their users’ requirements and constraints, which will require inventing new scalable approaches to training and building these systems. To achieve my research agenda, I aim to lead a lab that advances the foundations for building large-scale AI systems and **openly shares findings and artifacts, including datasets, training recipes, and model weights, with the community.** My research will pursue three connected directions: (1) automating post-training to enable systems to iteratively improve while remaining aligned with users’ intents, (2) developing a science of data-efficient pretraining and explaining how capabilities emerge during training, and (3) designing flexible foundation models that adapt their capacity and computation to diverse deployment needs.

Automating Post-Training A key ingredient of self-improving systems is the ability to learn from experience and adapt to new tasks and environments. I am interested in building agentic systems that continually improve their capabilities while remaining aligned with users’ intents. As these systems become more autonomous, their ability to revise strategies, use tools, and collaborate with other agents creates new opportunities for improvement, but also new pathways for misaligned behavior. **How can we automate post-training so that increasing capabilities remain grounded in users’ goals and constraints?** I will lead efforts to build reliable recipes for post-training these frontier systems that select and synthesize training data while accounting for the available resources. This effort will require creating new synthetic training datasets aligned to the target tasks, as well as new learning algorithms that teach agents to coordinate, recover from failures, and improve collective performance. Finally, I will lead efforts to build new tools to accurately evaluate their capabilities and monitor misalignment in multi-agent systems. My **prior work on efficient data selection** [13], **synthetic data generation** [12], and **auditing the capabilities of foundation models** [9], together with my **ongoing work on learning to generate targeted synthetic data** to improve downstream performance, will provide a strong foundation for this effort.

Science of Pretraining Foundation Models The scarcity of high-quality pretraining tokens is now a critical bottleneck for continued progress in foundation models. While scaling laws guide the allocation of model size, training tokens, and compute, they do not address how to achieve further performance gains under this practical constraint. In particular, we do not fully understand how different domains, languages, and modalities interact during pretraining to support or hinder particular capabilities. I will lead efforts to develop a **science of pretraining** that makes data efficiency central, identifies principles for achieving meaningful performance gains under limited token budgets, and explains how capabilities emerge during training. My research on studying **pretraining under constrained resources** [6] and **pretraining under finite token budgets** [8], together with my **ongoing work on how capabilities emerge during pretraining**, provides a strong foundation for these efforts.

Designing Flexible Foundation Models Today’s foundation models offer limited configuration options, creating challenges for practitioners deploying them across heterogeneous settings. Deployment requires balancing electricity use, latency, task performance, and hardware availability. I will lead efforts to develop simple, cost-efficient methods for adapting foundation models to these competing requirements. I envision **models that can be rapidly specialized or reconfigured, dynamically adjusting their architecture and inference-time computation** to match task complexity and available resources. This line of work will explore model plasticity, adaptive computation through looped transformers, and hybrid architectures combining attention with efficient state-space components. The goal is to *maximize intelligence per watt*, enabling models to scale their capacity and computation up or down while delivering accurate results efficiently. My **prior work on model size interpolation for pretrained and post-trained models, which supports scaling in model parameter space** [4, 5, 17], along with my ongoing work on model plasticity, will provide a strong foundation for these efforts.

References to My Work

- [1] T. Abdullahi, I. Gemou, **N. V. Nayak**, G. Murtaza, S. H. Bach, C. Eickhoff, and R. Singh. K-paths: Reasoning over graph paths for drug repurposing and drug interaction prediction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.
- [2] S. H. Bach*, V. Sanh*, Z.-X. Yong, A. Webson, C. Raffel, **N. V. Nayak**, A. Sharma, T. Kim, M. S. Bari, T. Fevry, Z. Alyafeai, M. Dey, A. Santilli, Z. Sun, S. Ben-David, C. Xu, G. Chhablani, H. Wang, J. A. Fries, M. S. Al-shaibani, S. Sharma, U. Thakker, K. Almubarak, X. Tang, D. Radev, M. T.-J. Jiang, and A. M. Rush. PromptSource: An integrated development environment and repository for natural language prompts. In *Meeting of the Association for Computational Linguistics (ACL) Demonstration*, 2022.
- [3] BigScience Workshop, including **N. V. Nayak**. BLOOM: A 176b-parameter open-access multilingual language model. *Journal of Machine Learning Research*, 25(422):1–74, 2024.
- [4] S. Kangaslahti, J. Geuter*, **N. V. Nayak***, M. Fumero*, F. Locatello, and D. Alvarez-Melis. Understanding layer patching in model size interpolation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2026.
- [5] S. Kangaslahti, **N. V. Nayak***, J. Geuter*, M. Fumero, F. Locatello, and D. Alvarez-Melis. Boomerang distillation enables zero-shot model size interpolation. In *The Fourteenth International Conference on Learning Representations*, 2026.
- [6] A. Khandelwal, T. Yun, **N. V. Nayak**, J. Merullo, S. H. Bach, C. Sun, and E. Pavlick. \$100K or 100 Days: Trade-offs When Pre-Training with Academic Resources. In *Conference on Language Modeling (COLM)*, 2025.
- [7] Y. Kordi, **N. V. Nayak**, M. Zuo, I. Nguyen, and S. Bach. Revisiting generalization across difficulty levels: It’s not so easy. In *19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026.
- [8] J. Lakha, **N. V. Nayak***, B. Liu*, S. M. Kakade[†], and D. Alvarez-Melis[†]. Self-Distillation for Data-Scarce Language Model Pretraining. In *High-Dimensional Learning Dynamics (HiLD) @ ICML 2026*, 2026.
- [9] M. Lewis*, **N. V. Nayak***, P. Yu, Q. Yu, J. Merullo, S. H. Bach, and E. Pavlick. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2024.
- [10] N. V. Nayak. *Zero-Shot Generalization with Structured Knowledge, Composition, and Synthetic Datasets*. PhD thesis, Brown University, 2025.
- [11] **N. V. Nayak** and S. H. Bach. Zero-shot learning with common sense knowledge graphs. *Transactions on Machine Learning Research*, 2022.
- [12] **N. V. Nayak**, Y. Nan, A. Trost, and S. H. Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
- [13] **N. V. Nayak**, P. Rodriguez-Diaz, N. Hulkund, S. Beery, and D. Alvarez-Melis. A critical look at targeted instruction selection: Disentangling what matters (and what doesn’t). In *International Conference on Machine Learning*, 2026.
- [14] **N. V. Nayak***, P. Yu*, and S. H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations (ICLR)*, 2023.
- [15] W. Piriyaikulij, C. Menghini, R. Briden, **N. V. Nayak**, J. Zhu, E. Raisi, and S. H. Bach. TAGLETS: A system for automatic semi-supervised learning with auxiliary data. In *Conference on Machine Learning and Systems (MLSys)*, 2022.
- [16] V. Sanh*, A. Webson*, C. Raffel*, S. H. Bach*, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, **N. V. Nayak**, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, and A. M. Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*, 2022.

- [17] Y. Zhou, S. Kangaslahti*, J. Geuter*, **N. V. Nayak**, M. Fumero, F. Locatello, and D. Alvarez-Melis. Thinking at the right size: Amortized distillation across post-trained llms. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

Other References

- [18] S. Biderman, M. A. Khan, N. Mireshghallah, C. Arnett, F. Barez, and N. Saphra. Position: Don’t just “fix it in post”: A science of AI must study training dynamics. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.
- [19] D. Blalock. 2024-4-28 arxiv roundup: Data and scaling, backlog highlights part 3. Davis Summarizes Papers, Apr. 2024. Includes coverage of “Learning to Generate Instruction Tuning Datasets for Zero-Shot Task Adaptation”.
- [20] K. Clark and P. Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36:58921–58937, 2023.
- [21] H. Kudiabor. AI’s computing gap: Academics lack access to powerful chips needed for research. *Nature*, 636: 16–17, Nov. 2024. doi: 10.1038/d41586-024-03792-6. News article covering Khandelwal et al., “\$100K or 100 Days: Trade-offs when Pre-Training with Academic Resources”.
- [22] Y. Liu. How to generate instruction datasets from any documents for LLM fine-tuning. Towards Data Science, Mar. 2024. Article demonstrating Bonito for generating synthetic instruction-tuning datasets.
- [23] National Science Foundation. CAIG: Navigating the climate science deluge: Training language models to assist in comprehensive assessments, 2024. Award No. 2425380.
- [24] OpenAI. The Hugging Face incident and the road ahead, Aug. 2026. Published August 26, 2026. Accessed October 3, 2026.
- [25] Thoughtworks. Synthetic data for testing and training models. Thoughtworks Technology Radar, Volume 31, Oct. 2024. Technology Radar entry; Trial ring. Mentions Bonito as a tool for generating synthetic instruction-tuning data.
- [26] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6(2):171–178, 1996. doi: 10.1016/S0959-4388(96)80070-5.
- [27] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.