Building the Data Foundation for Open and Sustainable AI Development

Artificial intelligence (AI) has entered a transformative era marked by large, extensively trained models. These models, often referred to as foundation models, can perform a wide range of tasks and empower diverse applications. At the core of these models is the data they are trained on, which is commonly regarded as the secret sauce for frontier models and companies. Despite their critical role, current data practices for developing these models are unscientific: they are often opaque, unprincipled, costly, and even fraught with legal concerns, casting shadows over the sustainability of AI in the future. The goal of my research is to **build the data foundation to advance and sustain open research and development of AI**. This data foundation includes an extensive collection of data artifacts that can shape model training, backed by systematic thinking about data paradigms, data creation pipelines, data optimization algorithms, and the fair and responsible use of data (Fig. 1). To make progress toward this goal, I have led the following research efforts:

- **Data creation in the quest for generality (§1):** I study how to create and use data to make models generalize better. I am among the early pioneers of instruction tuning, which aims to make models generalize across tasks by following natural language instructions. In particular, I proposed a unified task representation schema and curated SuperNaturalInstructions [1]—the earliest large-scale instruction dataset with 1,600+ NLP tasks, demonstrating the key scaling factors for training instruction-following models. This dataset and the insights have been used in building early foundation models in industry (e.g., Google's FLAN [23], PaLM [24], and Meta's Llama [25]) and also enabled various explorations of more generalizable models, including my work on instructable embeddings [2] and task-adaptive neural architectures [3].
- **Leveraging synthetic data to accelerate AI (§2):** I study the pipelines and principles for using synthetic data to improve models. I introduced Self-Instruct [4]—the first pipeline for generating diverse tasks using language models (LMs) and improving the LMs by bootstrapping off their own generations. This pipeline leverages the generative nature of LMs and significantly lowers the cost of data creation, leading to wide adoption in building instruction-following models of different modalities (e.g., Alpaca [26], LLaVa [27]), improving foundation models (e.g., Llama 3 [28], Nemotron [29]), and specializing them in different scenarios (e.g., coding [30], tool use [31]). Furthermore, my recent work [5] introduced an algorithmic framework to combine the strengths of both synthetic and human data for better performance, opening the door to a hybrid data creation paradigm that is better quality while efficient.
- **Building fully open language models (§3):** Open science is critical for innovation, collaboration, and accountability, while most of today's foundation models are trending oppositely. I co-lead the post-training efforts within the OLMo (Open Language Models) project [6]. Together with the team, I have systematically explored instruction tuning [7], preference modeling [5], reinforcement learning from human feedback [8, 9], and how to integrate them to match frontier proprietary models [10]. The resulting Tülu series of models have been downloaded over 250K times, and OLMo post-trained models over 100K times. The artifacts, particularly the data, are all released with clear documentation. This has enabled research and education that relies on fully open models and training procedures, including some of mine [11–13]. The first OLMo paper was also recognized with the Best Theme Paper Award at ACL 2024.

As the research and application landscape of AI rapidly expands, I believe that data forms the foundation for driving innovation, fostering collaboration, and achieving many desiderata of AI (e.g., safety, clear license). My future research plans to tackle emerging problems centered around data, with emphasis on 1) algorithms that optimize and expand data at scale, 2) mechanisms for the fair and responsible use of data, and 3) data for real-world interactions involving many modalities (§4). These efforts will be conducted openly in collaboration with the growing open community, extending transparent access to AI and the benefits of AI to a wider public.

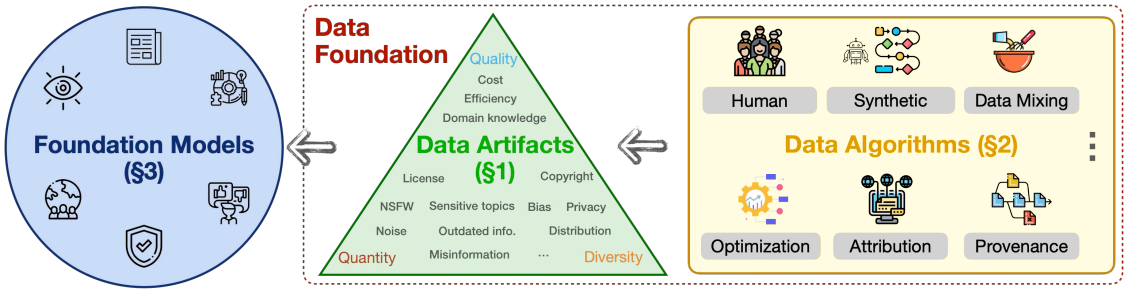

Fig. 1: Data plays a foundational role in the development, understanding, and oversight of AI. While extensive attention is focused on models, a sustainable future for AI calls for resolving key challenges in data and innovating on data algorithms to build a scientific data foundation. My research spans data algorithms, data artifacts, and training foundation models.

1 Data Creation in the Quest for Generality

The major focus of AI is progressing from single-purpose applications to a new stage aiming to build better general-purpose models. This shift toward generality is not just driven by improving model performance; it is also essential for creating more advanced systems that require broad capabilities to handle complex tasks. My research advances the study of generality through 1) creating datasets with a unified view of tasks, 2) expanding studies on models' key capabilities, and 3) optimizing data to achieve the right balance for better generalization. This has led to the early work on instruction tuning, a step that is nowadays ubiquitous in building foundation models. I also co-organized the first workshop on instruction tuning at NeurIPS 2023.

Super-NaturalInstructions: pioneering instruction tuning. Instruction tuning trains models to better follow instructions to perform tasks. I began exploring this concept before ChatGPT was released [1]. Based on earlier work on a small set of NLP tasks, I proposed a unified schema to represent tasks, which consist of natural language task definitions, positive and negative demonstration examples, example explanations, and instances of input and output (Fig. 2). This unified schema enabled collaboration with 88 NLP practitioners globally to collect 1,676 commonly used NLP tasks and annotate detailed instructions and meta categories for them, named Super-NaturalInstructions (SuperNI). SuperNI made it possible to conduct rigorous experiments, demonstrating that models' task-level generalization performance increases log-linearly as we scale up the task diversity and model size, and the information included in instruction also matters. This provided the early empirical basis for instruction tuning and building general-purpose models. The dataset was also used by companies such as Google and Meta for building their early foundation models [23–25]. SuperNI has opened up explorations in building various models that use instructions to generalize better. For example, we built one of the earliest instruction-following embedding methods and achieved sota performance on out-of-distribution retrieval [2], and we showed that a hyper-network can use instructions to generate adaptive model weights [3].

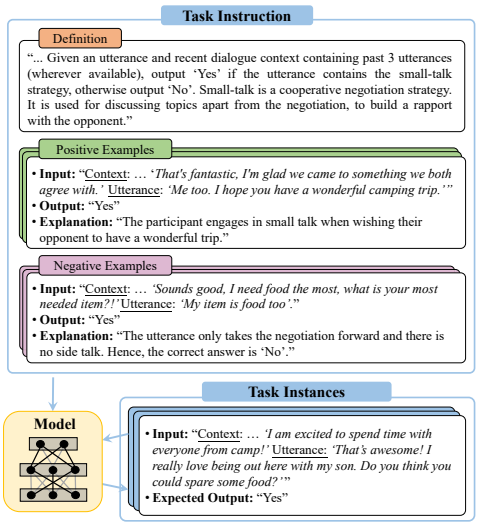

Fig. 2: An example task from SuperNI, where the model is expected to perform the task instances by following the task instruction.

Specialized datasets for expanding models' key capabilities. The quest for generality does not contradict the need for specialized datasets. In fact, specialized datasets still drive the exploration of the model's key capabilities and potential use cases. My research has contributed to the development of several widely used datasets for specialized purposes. For instance, DROP [14] focused on the phenomenon of discrete reasoning over information contained in passages. MultiModalQA [15] extended the traditional question-answering paradigm to multi-modal by incorporating textual, tabular, and visual data. The lay language summarization dataset of biomedical scientific reviews [16] helped improve health literacy. More recently, TurkingBench [17] used turking interfaces to benchmark web agents. Each of these datasets has explored the frontier of model capabilities, enabling them to address increasingly complex and diverse tasks.

2 Leveraging Synthetic Data to Accelerate AI

Despite the history of data-driven AI progress, annotating datasets has long posed challenges for researchers and practitioners in the field—challenges I experienced firsthand when building the datasets mentioned above. My research studies how to leverage synthetic data that is generated by models to tackle these challenges and even outperform human annotations. I started the first work in using language models to generate diverse and creative tasks in Self-Instruct [4]. This was widely adopted and expanded upon for building foundation models, such as the popular Stanford Alpaca [26], Mata's Llama-3 [28] and Nvidia's Nemotron [29]. The findings in Self-Instruct also popularized the use of synthetic data in AI research, including works that I mentored or collaborated on [18–20]. Today, as models keep improving and outperform human experts in many cases, synthetic data becomes even more critical in achieving model self-improvement or weak-to-strong alignment, while risks such as model collapse [32] and copyright [33] are also raised. Research in this area is still in full swing. I am continuing several projects and co-organizing a tutorial on "synthetic data in the era of LLMs."

Self-Instruct: the promise of synthesizing diverse tasks with models. Findings from SuperNI suggest that diversity of training tasks is critical for instruction tuning. Early studies, including early ChatGPT [34], rely

on human-written tasks that are both costly and limited in creativity. For example, they are often short and focused on typical NLP tasks (e.g., question answering). In Self-Instruct [4], I found that pre-trained models, because they have seen many texts on the web, are capable of generating tasks that represent broader and more complex human needs (e.g., filling out an application form). Based on this, I devised a pipeline to scale up the generation process (Fig. 3), which 1) starts off with a small pool of tasks, 2) samples examples from the pool to prompt LMs to generate new tasks and 3) corresponding instances, 4) filters invalid generations, and 5) expands the task pool. Iterating this process ends up bootstrapping a large number of diverse tasks that can be used to train models. Notably, the initial Self-Instruct assumes only a vanilla pre-trained model is available, and the model is primarily self-improving to follow instructions. Self-Instruct was quickly adopted by both academia and industry after its release and significantly inspired two lines of research: 1) distilling data from frontier models to train smaller models, which leads to the prosperity of the open community on finetuning LMs for different purposes; 2) improving frontier models by encouraging them to self-improve without accessing other models, which is a key focus for LM research nowadays and is believed to have scaling potential to change the current training paradigms.

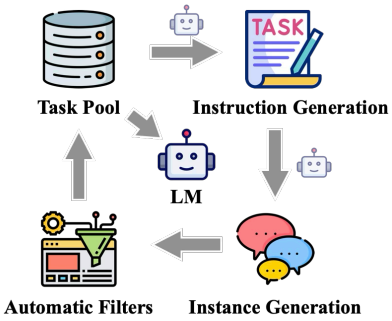

Fig. 3: Self-Instruct uses an LM to bootstrap diverse tasks and train the LM.

Hybrid Preferences: learning where we do and do not need humans. Synthetic data is appealing and easy to scale, but it is also prone to various errors and biases. Therefore, humans should still be consulted when necessary to improve data quality. My recent work [5] studies the trade-off between human vs. model annotations for preference feedback. Surprisingly, we found that naively using human-annotated labels for all instances results in lower performance compared to using synthetic labels by GPT4, suggesting that careful routing between the two is needed. We formulate this as an optimization problem to maximize task performance by finding an optimal routing strategy for annotating each instance by either humans or models. This data optimization framework not only improves the performance of several representative preference datasets but also provides categories where human annotators still win (e.g., instances with moderate intent complexity). It also has the potential to guide other data decisions and optimize data at scale, as I will discuss in §4. I believe that hybrid mode is the right way to use human expert annotations in the future, particularly for complex tasks where human annotations are expensive, and we are only at the beginning of this new data paradigm.

3 Building Fully Open Language Models

Foundation models hold the promise for the future of AI and many applications impacting our society. However, the current foundation models are mainly developed by for-profit companies with little incentive to cast a scientific light on their products. This closed nature hinders understanding of models' inner workings and creates challenges for researchers, policymakers, and the public to assess their efficacy, safety, privacy, or legal implications. Later in my PhD, I had the opportunity to participate in the OLMo project [6] to build fully open LMs, where I co-led the post-training efforts. We have built two series of models: Tülu, which builds upon base models from the community (e.g., Llama) but without full information about pre-training, and OLMo, which starts from pre-training and provides transparency to the full LM training pipeline. These models can match and even outperform frontier proprietary models of their sizes. You can interact with them in this Playground. We have created and released all artifacts for developing these models, particularly the data that no frontier companies are willing to share, with clear documentation and code for reproducing them. This not only enables research and education that rely on fully open models or their training procedures, but also demonstrates the possibility of developing foundation models openly to serve the public interest.

Tülu and OLMo: fully open recipes for language model post-training. The training of large LMs like ChatGPT has converged into broadly two stages: pre-training on extensive text to grasp knowledge and skills, and post-training with various techniques to align with human intents and further improve. The post-training side, because of the complexity of training methods and many direct use cases, attracts significant attention, leading to many pieces of innovations, including datasets, algorithms, and evaluations. However, it is unclear what really excel, how to integrate them, and how they truly compare with proprietary models that are heavily engineered. Therefore, I initiated and co-led the Tülu series of work, which starts from open-weight pre-training models such as Llama and systematically studies the integration of post-training techniques. Tülu v1 [7] that I led studied instruction tuning datasets, effects for mixing them, relation to model sizes, and systematic evaluations for the post-trained models. Later, I co-led Tülu v2 [8] that incorporated preference training with DPO, and Tülu v2.5 [9] that systematically explored RLHF training algorithms and datasets. Recently, I have advised a big team at AI2 on releasing Tülu v3 [10] that integrates the latest and best techniques to match frontier mod-

els. Efforts in Tülu also guided the pre-training design of OLMo and were all applied to build OLMo-instruct [6], providing a fully open recipe from pre-training to post-training for studying the science of LMs.

Better understanding of LMs. Besides the training of LMs, I also led research in understanding how LMs work in different settings. This includes: 1) math reasoning [21], for which my work was the first to show the numeracy encoded in pre-trained models and its failure in extrapolation; 2) factual knowledge [18, 19], for which I mentored students to study LMs' hallucination types and their factuality on time-sensitive questions; 3) long-context capability [13], for which my recent work revealed the existence of dedicated long-range attention heads explaining for its success; and 4) learning curriculum [11], for which my early work demonstrated clearly different learning patterns for different types of knowledge in pre-training. These findings has led a large body of follow-up work and serve as the empirical bases for further development of LMs.

4 Future Directions

I envision a future where AI systems not only advance in capability and generality but also operate within an ecosystem that is legally sound, safe, open, and beneficial to all. My future work will focus on enhancing and expanding the data foundation to achieve these objectives. Below are key research directions I plan to pursue.

Data optimization and expansion at scale. As the link between data and model performance becomes increasingly evident and datasets grow in scale and complexity, research on automated data optimization is urgently needed. Studies like my early work [22] have been conducted on data selection strategies, which primarily prune existing datasets. However, most methods are still ad hoc or only consider simple dataset characteristics when applied at a large scale, leaving room for improvement. Another even more important data question when building general-purpose models is, "where should we expand new data?" Current practices are usually driven by human intuitions, interests, or limited empirical results, without a principled approach to guide data expansion. Moreover, as the target capabilities become more complex and the models also grow stronger and harder to supervise, we should also decide "where to source data", as many sources, including human annotators, are prone to biases and errors. All these decisions form a large optimization space for data, and I am excited to establish a better formulation of this problem and propose general methods to optimize data for better performance in different stages of training, building upon my earlier work [5].

Fair and responsible use of data. Many pressing issues around privacy, copyright, and ethical integrity exist in building current foundation models, especially in terms of AI practitioners scraping content people share on the web without notifying and crediting them properly. However, given the diverse sources of datasets and the fusion nature of generative models, ensuring the fair and responsible use of data is extremely challenging. I plan to work on two lines of research toward this goal. First, we need better attribution from model outputs to the training data, which will provide a better tool for people to investigate the use of their data and ideally get fair credits for their data contributions. Existing studies in influence functions and watermarking provide good starting bases but need more work to make them scalable and reliable. Second, we need better mechanisms for tracking the provenance of data used in model training to increase data legitimacy and avoid data contamination. Researchers in the research community, including myself, have initiated efforts toward this goal. For example, in building Tülu 3 [10], I manually track the provenance and licenses of datasets, leading to a clean while performant instruction tuning dataset that people can use, even commercially. However, this is not scalable to a large set of sources, and the tracking starts failing when datasets are transformed in many ways. I believe this takes joint efforts from computer science, law, and policy studies. Building upon artifacts released in OLMo and Tülu, especially the training data, I am excited to study both directions and their interplay with model training with interdisciplinary collaborations.

Real-world interactions involving many modalities. I believe that AI systems will keep improving with a future of helping people in the physical world. One key challenge for building such systems is to collect data that involves real-world interactions. Such data is sparse on the web, and collecting it raises many concerns, such as privacy and safety. I am interested in expanding my research in this direction, particularly from the data perspective. My past research can help advance this in multiple ways. The unified view of tasks can help formulate a new paradigm of tasks incorporating multi-modal inputs (e.g., sensory data) and real-world interactions (e.g., robot execution in an environment). The experience with generative synthetic data can provide insights into using generative models to simulate interactions at scale and produce diverse data. The techniques in building open LMs can also guide the training for models of multi-modalities beyond language and vision. I view this as my long-term research direction—it is a continuation of my quest toward better generality after the success of language foundation models, although I understand the complexity of potentially combining multiple subjects, including language, perception, robotics, planning, etc. I am looking forward to collaborating with experts in these fields to move forward.

References to my work

[1] **Y. Wang***, S. Mishra*, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, *et al.*, "Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks," in *EMNLP*, 2022.

[2] H. Su, W. Shi, J. Kasai, **Y. Wang**, Y. Hu, *et al.*, "One embedder, any task: Instruction-finetuned text embeddings," in *ACL Findings*, 2023.

[3] H. Ivison, A. Bhagia, **Y. Wang**, H. Hajishirzi, and M. Peters, "HINT: Hypernetwork instruction tuning for efficient zero-and few-shot generalisation," in *ACL*, 2023.

[4] **Y. Wang**, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, *et al.*, "Self-Instruct: Aligning language models with self-generated instructions," in *ACL*, 2023.

[5] L. Miranda*, **Y. Wang***, Y. Elazar, S. Kumar, V. Pyatkin, *et al.*, "Hybrid preferences: Learning to route instances for human vs. AI feedback," *arXiv*, 2024.

[6] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. Jha, H. Ivison, I. Magnusson, **Y. Wang**, *et al.*, "OLMo: Accelerating the science of language models," in *ACL*, 2024.

[7] **Y. Wang***, H. Ivison*, P. Dasigi, J. Hessel, T. Khot, *et al.*, "How far can camels go? exploring the state of instruction tuning on open resources," in *NeurIPS*, 2023.

[8] H. Ivison*, **Y. Wang***, V. Pyatkin, N. Lambert, M. Peters, *et al.*, "Camels in a changing climate: Enhancing LM adaptation with Tülu 2," *arXiv*, 2023.

[9] H. Ivison, **Y. Wang**, J. Liu, Z. Wu, V. Pyatkin, *et al.*, "Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback," in *NeurIPS*, 2024.

[10] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, *et al.*, "Tülu 3: Pushing frontiers in open language model post-training," *arXiv*, 2024.

[11] Z. Liu*, **Y. Wang***, J. Kasai, H. Hajishirzi, and N. A. Smith, "Probing across time: What does RoBERTa know and when?" In *EMNLP Findings*, 2021.

[12] A. Liu, X. Han, **Y. Wang**, Y. Tsvetkov, Y. Choi, *et al.*, "Tuning language models by proxy," in *COLM*, 2024.

[13] W. Wu, **Y. Wang**, G. Xiao, H. Peng, and Y. Fu, "Retrieval head mechanistically explains long-context factuality," *arXiv*, 2024.

[14] D. Dua, **Y. Wang**, P. Dasigi, G. Stanovsky, S. Singh, *et al.*, "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs," in *NAACL-HLT*, 2019.

[15] A. Talmor, O. Yoran, A. Catav, D. Lahav, **Y. Wang**, *et al.*, "MultiModalQA: Complex question answering over text, tables, and images," in *ICLR*, 2021.

[16] Y. Guo, W. Qiu, **Y. Wang**, and T. Cohen, "Automated lay language summarization of biomedical scientific reviews," in *AAAI*, 2021.

[17] K. Xu, Y. Kordi, K. Sanders, **Y. Wang**, A. Byerly, *et al.*, "Tur[k]ingBench: A challenge benchmark for web agents," *arXiv*, 2024.

[18] B. Zhao*, Z. Brumbaugh*, **Y. Wang***, H. Hajishirzi, and N. A. Smith, "Set the clock: Temporal alignment of pretrained language models," in *ACL Findings*, 2024.

[19] A. Mishra, A. Asai, V. Balachandran, **Y. Wang**, G. Neubig, *et al.*, "Fine-grained hallucination detection and editing for language models," in *COLM*, 2024.

[20] A. Asai, Z. Wu, **Y. Wang**, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *ICLR*, 2024.

[21] E. Wallace*, **Y. Wang***, S. Li, S. Singh, and M. Gardner, "Do NLP models know numbers? probing numeracy in embeddings," in *EMNLP-IJCNLP*, 2019.

[22] S. Swayamdipta, R. Schwartz, N. Lourie, **Y. Wang**, H. Hajishirzi, *et al.*, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *EMNLP*, 2020.

References to other work

[23] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, *et al.*, "The FLAN collection: Designing data and methods for effective instruction tuning," in *ICML*, 2023.

[24] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, *et al.*, "PaLM 2 technical report," *arXiv*, 2023.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[26] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, *et al.*, "Stanford Alpaca: An instruction-following Llama model," *GitHub repository*, 2023.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[28] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, *et al.*, "The Llama 3 herd of models," *arXiv*, 2024.

[29] B. Adler, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, *et al.*, "Nemotron-4 340B technical report," *arXiv*, 2024.

[30] Y. Wang, H. Le, A. Gotmare, N. Bui, J. Li, *et al.*, "CodeT5+: Open code large language models for code understanding and generation," in *EMNLP*, 2023.

[31] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, *et al.*, "ToolLLM: Facilitating large language models to master 16000+ real-world APIs," in *ICLR*, 2024.

[32] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, *et al.*, "AI models collapse when trained on recursively generated data," *Nature*, 2024.

[33] M. S. Gal and O. Lynskey, "Synthetic data: Legal implications of the data-generation revolution," *Iowa Law Review*, 2023.

[34] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, *et al.*, "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.