I envision the democratization of natural language processing (NLP) and artificial intelligence (AI) knowledge, where people with diverse skill levels and research backgrounds can: build, use, analyze, and evaluate models; collaborate to solve research problems; and accelerate advances in NLP and AI. AI and NLP have made remarkable progress from recent, large-scale training on massive datasets. These technologies are being developed and used by many cross-disciplinary researchers and practitioners. People with scant computer science training—including physicians, translators, and historians—now rely on AI models for work problems that can be solved by using massive amounts of data.

I focus on ways to make AI/NLP more accessible to researchers, practitioners, and users. How can we encourage model builders and practitioners to work as a community to broaden the appeal and utility of NLP and AI models across disciplines? How can we make it easier for them to formulate and answer complex real-world questions using these technologies and ensure these models are robustly evaluated? How can we guarantee that global citizens and English speakers enjoy the benefits of NLP and AI advances in equal measure?

During my Ph.D., I made the following contributions towards realizing this long-term vision.

- **Transparent and Reliable Evaluation Methodologies for NLP (§1).** I designed methodologies and interfaces to make model evaluations more transparent, consistent, and reliable [1, 2, 3, 4, 5].
- **Flexible and Customizable Inference Algorithms (§2).** I built algorithms for flexible and customizable language generation in the areas of collaborative inference between diverse models [6]; controlled generation [7]; and fast, accurate inference [8, 9]. All of these methods avoid the computationally (and thus financially and environmentally) expensive training process of large models.
- **Efficient and Accessible NLP (§3).** I introduced and empirically demonstrated efficient architectures and learning paradigms for state-of-the-art NLP models [10, 11, 12, 13, 14]. More efficient methods lower the cost of developing and using these models, making them deployable to less-well-funded fields or institutions.

## 1 Transparent and Reliable Evaluation Methodologies for NLP

Evaluation is the tool that lets us measure progress, understand model behaviors or failures, and guide future research directions. Evaluating NLP systems presents a serious challenge, particularly for language generation tasks such as machine translation and summarization. Language generation is inherently open-ended, and generation quality cannot be measured using a simple metric like classification accuracy on ImageNet. A large body of NLP work therefore develops automatic evaluation methods that enable fast, inexpensive development cycles [15]. However, this progress in evaluation has been largely overlooked by researchers focused on model advancements [16, 17] mainly because they prioritize consistency of evaluation practices over time. I believe that this separation between generation and evaluation offers an opportunity for each sub-community to more rapidly benefit from the advances of the other.

Bidimensional leaderboards: an interface that bridges modeling and evaluation research. My NAACL 2022 work [1] introduces an
abstraction of leaderboards, called **bidimensional leaderboards** (BILLBOARDS), that simultaneously facilitates progress in natural language generation and its evaluation (Fig. 1).¹ A BILLBOARD accepts two types of submissions related to a given task and dataset: **generators** and **metrics**. Unlike conventional leaderboards, BILLBOARDS do not tie model ranking to a predetermined set of metrics: generators are ranked based on the most reliable metric currently available. Metric submissions are ranked by their correlations to human judgments, and each is stored as an executable program, which is then used to evaluate future generation submissions. This interface facilitates communication between the modeling and evaluation sub-communities. I currently maintain four BILLBOARDS over machine translation, summarization, and image captioning tasks. Each has more than 15 metrics to date, promoting changes in how evaluation is performed and studied in these sub-areas. BILLBOARD’s built-in analysis shows that **most automatic evaluations overrate machine over human-written generation**, demonstrating the importance of updating metrics as generation models become stronger (and perhaps more similar to humans) in the future. Though this work was published as recently as July 2022, it has already impacted evaluation practices of other tasks in NLP and beyond [18, 19], and I anticipate it will continue to do so.

**Transparent and scalable human evaluation.** As AI technology becomes further enmeshed in our society, I believe it is vital to systematically collect feedback about models from human users. For example, image captioning technology is already used to improve information accessibility for people with visual impairments; my NAACL 2022 work establishes a transparent human evaluation protocol for state-of-the-art image captioning models [2]. Unlike evaluations conducted by previous work (e.g., [20, 21]), our human-in-the-loop evaluation demonstrates that machine-generated captions continue to fall short of human-written ones. Our follow-up work [4] uses crowdsourcing to further scale up human-in-the-loop evaluation to more NLP tasks and models. GENIE is the first leaderboard known to support human-in-the-loop evaluation for a broad set of natural language tasks.² It has received 63 submissions over six benchmarks, including summarization, machine translation, and commonsense reasoning tasks. We used GENIE with WMT [22], an annual and long-standing benchmarking effort for machine translation and NLP, to evaluate system submissions from more than 20 institutions in academia and industry.

**Future directions.** As my work illustrates, continual, community-wide efforts are needed to evaluate AI systems. I am excited to extend my effort beyond NLP to speech and computer vision. Moreover, I would like to extend BILLBOARDS to **multidimensional leaderboards**. AI models have many more assessment dimensions than output quality (on a particular test set), such as training and inference efficiency, sample efficiency, environmental impacts, and robustness. These dimensions, often ignored in the current evaluation paradigm, are nonetheless critical to better serving practitioners’ needs [23, 24, 25]. I plan to build evaluation platforms where users can explore trade-offs of these different aspects to find a model that best fits their needs and where researchers can contribute AI models that have different strengths. I believe that evaluation should be aligned with the needs of a wide range of stakeholders.

## 2 Flexible and Customizable Inference Algorithms

Modern AI and NLP models undergo two stages: training, where model parameters are learned based on large datasets, and **inference**, where the model is used to generate desired outputs. Inference is thus a core algorithmic component of NLP systems that has been actively

---


² [https://genie.apps.allenai.org/](https://genie.apps.allenai.org/).
studied in NLP research. Generation of language handles complex structure and exponentially many possibilities; for a given output length $N$ and vocabulary size $V$, there are $V^N$ possibilities, making it impossible to enumerate and find the best one. Several inference algorithms (e.g., greedy/beam search) have proven effective on a variety of language generation tasks and continue to be used in recent large-scale models, such as GPT-3 [26] and Google Translate [27]. Overcoming the limitations of commonly used algorithms, I develop inference methods that enable flexible and customizable generation schemes: collaborative generation between two diverse generation models [6], efficient generation with parallel computation [9], and controllable generation [7].

**Best of both generators.** Combining diverse models (possibly from different institutions) can lead to further progress by leveraging their complementary strengths. Conventional ensembling methods make strong assumptions about vocabulary, tokenization, and probability factorization that often do not hold in practice. For example, the word “COVID-19” did not exist in the English language before 2020; when we combine models trained before and after 2020, we need to resolve divergent language generations from such differences. Similarly, these assumptions make it challenging to combine models with different specializations (e.g., medical/legal domains). I introduced the Twist algorithm, a simple yet effective method that relaxes these assumptions while avoiding any additional, expensive training. **Twist** performs standard beam search alternatively between two diverse generators with mutual guidance. Figure 2 shows results from combining a domain-specific model (medicine/law) with a generic model for German-to-English translation. Quality (y-axis) is measured using a state-of-the-art evaluation method from my BILLOARD [1]. Here, the number of domain training examples (x-axis) is much smaller than the generic training data (more than 30 million) since the creation of domain data requires bilingual speakers with domain expertise. **Even though the domain model performs poorly by itself, it improves the generic model through use of the Twist algorithm, demonstrating that Twist can make crucial use of complementary strengths.**

**Fast, parallel inference algorithm.** Inference speed is key to many AI applications, ranging from simultaneous machine interpretation to large image generation. Inference is conventionally done sequentially and word by word (or pixel by pixel), limiting the usage of fast parallel computations from modern hardware (e.g., GPUs and TPs). I anticipate that a major driving force of computing improvement will be increased parallelism, given physical constraints on the clock speed (i.e., the heat wall, [28]). For this reason, I believe that developing a parallelizable inference algorithm is one promising approach to fast generation. In my ICML 2019 work [9], I introduced a parallel algorithm that achieves similar generation quality with a speedup of more than 3x on a modern GPU compared to standard sequential inference. **This work has received continuing interest from the NLP, machine learning, and even speech communities (e.g., [29, 30, 31, 32]).**

**Future directions.** AI model usability depends heavily on inference algorithms. Developing flexible, customizable, and efficient inference algorithms is therefore essential to my long-term goal of building accessible NLP. Looking forward, I believe that there is a critical limitation of current algorithms that should be addressed to this end: current inference methods lack...
control over generation outputs, occasionally resulting in hallucinations and non-inclusive or toxic language. These problems can have serious implications for downstream applications, as observed in my evaluation work on image captioning [2]. My collaborators and I developed an A* search-based algorithm that enables lexical constraints on language generation [7], which won the best paper award at NAACL 2022. Building upon this work, I aim for the even more challenging controllable generation that, for instance, avoids toxic language (and images). Specifically, I plan to integrate my work on evaluation methodologies with an inference algorithm so that, for example, an evaluation metric from BILLBOARDS can capture the inclusiveness of generated language, which sends a useful signal to inference.

3 Efficient and Accessible NLP

The growing computational (and thus financial and environmental [23]) cost of large-scale AI/NLP models makes it difficult for many researchers to develop or even just use them. Efficient and accessible methods are needed to promote the accessibility of our technology and to ensure that people from diverse backgrounds can contribute to its progress from their various perspectives. I strongly believe that the inclusiveness of the AI community will be key to its success.

Finetuning off-the-self transformers into efficient RNNs. The transformer architecture [33] is a backbone of recent advances in NLP, computer vision, speech, computational biology, and beyond (e.g., [26, 34, 35]). Transformers outperform recurrent neural networks (RNNs) at the expense of their increased computational cost: their time and memory complexity scales quadratically with sequence length, in contrast to the linear complexity of RNNs. This computational requirement limits the usability of many strong transformers in the AI community. My EMNLP 2021 work [10] introduced transformer-to-RNN (T2R), a method that converts any off-the-shelf transformer into an efficient RNN counterpart by performing a small amount of finetuning (Figure 3). Drawing inspiration from the classical kernel methods in machine learning, T2R learns to approximate transformers’ quadratic computation during lightweight finetuning, resulting in a recurrent model with linear complexity. I empirically demonstrated that T2R generates long text with quality similar to the original transformer while achieving substantial speedup and memory savings (e.g., 10x speedup when producing 2048 consecutive words; Figure 4).

Future directions. Efficiency research is a subject of increasing interest to AI and NLP researchers. Developing efficient methods often requires a deep understanding of tasks and applications of interest. For example, I showed that a simple strategy of deep encoder, shal-
low decoder improves efficiency while retaining strong performance in sequence-to-sequence applications, such as machine translation [11]. I am passionate about developing efficient machine learning methods that are motivated by NLP applications. Conversely, I also find it exciting to draw mathematical connections among various approaches; such unified abstractions can help us better understand seemingly disparate research and inspire new advances, as illustrated by our ACL 2022 work [12]. Finally, efficiency research should not be limited to computational aspects. In particular, sample and annotation efficiency for data creation are also crucial to empower diverse AI applications. I had initial success in this attempt [36, 14], and I hope to expand such efforts.

4 Future Research Agenda

Dynamic and real-world evaluations. The AI train has already left the station: GPT-3 is used in more than 300 apps, and Google Translate translates more than 100 billion words per day. At present, most AI and NLP models are evaluated statically and only once, when papers are written. Since AI models have proven useful in many tasks, there has never been a more opportune time for the community to explore dynamic and real-world evaluations beyond fixed test data. My collaborators and I recently started RealTime QA, a dynamic benchmarking effort that evaluates question answering systems in real time [3]. Such dynamic evaluations benefit from interdisciplinary collaborations; real-time question answering systems can, for instance, facilitate emergency management of natural disasters and pandemics. I am excited to contribute to this effort from NLP and machine learning perspectives. I believe that real-world evaluations will reveal crucial challenges that guide our future research in many aspects. For example: How should a real-time system combat fake news or toxic content? How can we update an NLP system quickly and efficiently to fulfill real-time information needs? I am confident that my expertise in evaluation methodologies, efficient NLP, and language generation will add valuable insights to this research area.

Massively multilingual NLP. Current NLP data creation and model development focus heavily on the English language, leading to over-representation of English-centric problems (e.g., information needs about American politics). This lack of multilingual NLP research and resources limits the diversity and accessibility of AI technology and heightens the barriers to the use of many AI applications in the world. Indeed, English covers only one quarter of global web users;¹ I will pursue massively multilingual processing that benefits people around the world; I will contribute to creating linguistically diverse resources and advancing models that can be used for applications in many languages. During the early years of my Ph.D., my collaborators and I showed the possibility of effectively expanding AI models to diverse languages with smaller or even no labeled training data by using our multilingual vector representation [37, 38]. Multilingual vector representations continue to be studied in more recent work (e.g., [39]). I am particularly excited about exploring multilingual processing from the perspectives of inference algorithms and efficiency. Many global languages lack large knowledge sources, such as Wikipedia, but much current work assumes that user questions can be answered based solely on a knowledge source in the users’ own language [40, 41, 42]. Can we find an inference algorithm that systematically combines knowledge sources from various languages and provides answers to users regardless of their language? Can we develop models that efficiently process many languages? With the rich body of tools and methods currently available, I believe that the time is ripe to tackle these challenging research problems and improve the inclusiveness of language technologies for the billions of speakers of the 7,000+ languages other than English.

References


