Summary of Research and Contributions

Broadly, my research goal is to build machine intelligence that understands how the world works and interacts with humans safely and reliably. Specifically, I focus on enhancing the **commonsense reasoning** ability and **controllability** of neural language models as well as understanding their **fundamental limitations**.

The rise of large language models (LLM), such as ChatGPT, has led to an unprecedented amount of global attention—both excitements and concerns, in part due to the relatively limited understanding of machine intelligence. My research explores the potential limitations of large language models, specifically the fundamental constraints of transformers. Additionally, I investigate alternative, seemingly impossible pathways to develop machine intelligence beyond simply scaling up language models. My work highlights the importance of knowledge as well as training and inference time reasoning algorithms for the acquisition of commonsense reasoning and controllability. I demonstrate how smaller models developed in academia can still have an edge over larger industry-scale models, if powered with knowledge and/or reasoning algorithms. Concretely, my research branches around three themes:

- **Science of LLM.** Large language models demonstrate unprecedented capabilities, even noted as “sparks of AGI”. In stark contrast, the same models can still make basic errors, and struggle with simple, intuitive tasks. My work seeks to demystify large language models by exploring their fundamental limits through *compositional problems* that require strict multi-hop reasoning to derive correct predictions [1]. Additionally, I investigate the divergence in the configuration of machine and human intelligence by proposing and testing the *Generative AI Paradox* hypothesis [5].

- **Algorithm.** Language models, despite its scale or capability, still exhibit behaviors that are misaligned with user expectations. For example, generated text may contain offensive or toxic language, or fail to incorporate certain constraints user specified. To this end, my work investigate into *reinforcement learning algorithms* that unlearn undesirable behaviors [4] and *decoding time algorithms* that enforce faithful constraint satisfaction [3, 2], in order to achieve better controllability and enhance the safety and reliability of neural language models.

- **Knowledge.** Human-level language understanding grounds on a commonsense mental model of ‘how the world works’, which requires physical reasoning over objects and actions, along with higher-order event reasoning about complex situations. Today’s machines struggle with both. My research seeks to bridge this gap by enable machine to learn *multimodel script knowledge* from complex raw data, which leads to new SOTA performances on a dozen leaderboards that require grounded, temporal, and causal commonsense reasoning [6, 7].

Capabilities and Limits of Language Model. As large language models continue to make tangible real-world impacts, it is pressing to interpret their remarkable performance critically. My work Faith and Fate [1] takes a realistic look at the limitations of transformers in the context of compositional tasks. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that language mdoe solve compositional tasks by reducing multi-step compositional reasoning into linearized subgraph matching, without necessarily developing systematic problem-solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how autoregressive generations' performance can rapidly decay with increased task complexity. Beside understanding the fundamental limitations, I further explore the divergence in the configuration of machine and human intelligence by proposing and testing the Generative AI Paradox hypothesis: generative models, having been trained directly to reproduce expertlike outputs, acquire generative capabilities that are not contingent upon—and can therefore exceed——their ability to understand those same types of outputs [5]. This contrasts with humans, for whom basic understanding almost always precedes the ability to generate expert-level outputs. We show that although models can outperform humans in generation, they consistently fall short of human capabilities in measures of understanding, showing weaker correlation between generation and understanding performance, and more brittleness to adversarial inputs. Our findings support the hypothesis that models' generative capability may not be contingent upon understanding capability, and call for caution in interpreting artificial intelligence by analogy to human intelligence.

Constrained Decoding Algorithm. Conditional text generation often requires lexical constraints, i.e., which words should or shouldn't be included in the output text. While the dominant recipe for conditional text generation has been large-scale pretrained language models, prompted with or finetuned on the task-specific training data, such models do not learn to follow the underlying constraints reliably. In contrast, human could perform constrained generation out of the box without seeing any task specific examples. To enable such capability for neural language model, we propose NEUROLOGIC decoding [3], which effectively enforces the satisfaction of given lexical constraints by controlling the decoding stage of sequence generation. Neurologic decoding performs constrained optimization via beam-like search to find optimal sequences with respect to both likelihood and constraint satisfaction. Neurologic is powerful yet efficient. It handles any set of lexical constraints that is expressible under predicate logic, while its asymptotic runtime is equivalent to conventional beam search. I further built on this work through a new algorithm named NEUROLOGIC A*esque [2] inspired by the A* search algorithm, which incorporates heuristic estimates of future cost into the search procedure. We develop lookahead heuristics, which approximate cost of satisfying future constraints based on continuations of the sequence-so-far to aid Neurologic search. Perhaps surprisingly, we find that unsupervised models often match or outperform supervised approaches when powered with NEUROLOGIC, even when the latter is based on considerably larger networks. My works suggest the promise of inference-time algorithms to enable new capability of language models beyond scaling.

Reinforced Unlearning Algorithm. Large neural language models trained on an enormous amount of web text have excelled at numerous tasks. However, these same language models often exhibit undesirable behaviors, as they are usually trained to simply maximize the likelihood of their raw pre-training data. Undesirable behaviors are diverse and hard to avoid, control, or even specify *a priori*; I thus argue that it is critical to investigate ways to *unlearn* undesirable behaviors *post hoc*, while maintaining capacity for generating coherent and fluent language. We introduce Quantized Reward Konditioning (Quark), an algorithm for optimizing a reward function that quantifies an (un)wanted property, while not straying too far from the original model. Quark alternates between (i) collecting samples with the current language model, (ii) sorting them into quantiles based on reward, with each quantile identified by a reward token prepended to the language model's input, and (iii) using a standard language modeling loss on samples from each quantile conditioned on its reward token, while remaining nearby the original language model via a KL-divergence penalty. By conditioning on a high-reward token at generation time, the model generates text that exhibits less of the unwanted property. For unlearning toxicity, negative sentiment, and repetition, Quark outperforms both strong baselines and state-of-the-art reinforcement learning methods like PPO, while relying only on standard language modeling primitives.

Multimodel Script Knowledge. Over the last few years, many large-scale NLP and computer vision models have been trained on a combination of text, images, and manual annotations – yet, this approach has not been sufficient to 'solve' tasks like Visual Commonsense Reasoning (VCR), which requires grounded, temporal, and causal commonsense reasoning. My work introduces a new approach, where we train a model on multimodal and temporal data from YouTube [7]. We use new self supervised objectives to learn multimodal script knowledge. We dub our model MERLOT, short for Multimodal Event Representation Learning over Time. Our model sets new state of-the-art results on twelve video reasoning tasks, as well as on VCR. In doing so, it outperforms larger, industry-submitted models that that learn from static data: images annotated with object detections, and literal descriptions. We recently built on this work through a new model named MERLOT Reserve [6]. The idea is to learn connections between all modalities including sound to understand videos. Perhaps surprisingly, our work shows that integrating sound improves vision-and-text representations. We set a new state-of-the-art on VCR, even though it doesn't include any sound for models.

References

[1] Nouha Dziri*, **Ximing Lu***, Melanie Sclar*, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332. Curran Associates, Inc., 2023.

[2] **Ximing Lu**, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neuro-Logic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] **Ximing Lu**, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online, June 2021. Association for Computational Linguistics.

[4] **Ximing Lu**, Sean Welleck, Liwei Jiang, Jack Hessel, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[5] Peter West*, **Ximing Lu***, Nouha Dziri*, Faeze Brahman*, Linjie Li*, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The generative AI paradox: “what it can create, it may not understand”. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Rowan Zellers, Jiasen Lu, **Ximing Lu**, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16354–16366, 2022.

[7] Rowan Zellers*, **Ximing Lu***, Jack Hessel*, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.