I am interested in studying Natural Language Processing (NLP) systems which are interpretable, explainable and collaborate with humans in constructive and unique domains. Despite dramatic benchmark improvements with large scale, instruction fine-tuned LLMs, we lack explanations as to *why* our NLP systems produce certain outputs and *how* underlying abilities emerge, combine and interact to execute high-level tasks. I believe understanding such properties requires a systematic study of *behavior*, from the lens of (1) improved NLP evaluation, (2) new analytical tools motivated by our understanding of human cognition and (3) through the use of these insights to align behavior with human intent. Having moved between engineering-focused ML/NLP projects within the industry and research groups at Georgia Tech, I have a unique perspective on the dichotomy between our theoretical understanding of reasoning and abstraction in LLMs and the extent such abilities help, or fail, in improving human collaboration.

Localizing LM Behavior through Controllable Generation. Traditional evaluation (e.g., with task-specific automatic metrics) has become challenging due to the evolving failure modes in the shifting landscape of language model scale, design and dataset constructions. Current fine-grained analysis captures a subset of model errors, but a comprehensive study of *behavior* must catalogue far more than error identification. In my work at EMNLP, I realized one promising approach is to exhaustively model the full distribution of language operations through well structured and widely applied generation tasks like text revision. I proposed the first holistic fine-grained text-to-text evaluation by annotating every linguistic transformation performed by LLMs in a text simplification task [6]. Using this setup, annotators highlight spans to capture both successful *and* failed edits, and organize them into an expert-written typology of 21 types. By proposing a scheme grounded in linguistic analysis and annotating with a far richer signal in word-level quality, we were able to reveal fundamental differences between the distribution of simplification approaches performed by closed- and open-source LLMs and human writing which previous work could only observe through intuition or exemplars. From this, we designed the first reference-free automatic text simplification metric by training on a dual word and sentence level objective. I came to realize this approach is a powerful technique far beyond text simplification, and that lowering the replication barrier and encouraging standardization for span-based annotation studies is critical for guiding future work, which led to our EMNLP demo of a universal interface builder, Thresh, for fine-grained text evaluation [7]. Using Thresh, we further extended a standardized interface and data format to 12 other error evaluation studies across text generation tasks and has already streamlined ongoing evaluation projects at Georgia Tech and UT Austin. I am excited to further develop a unified fine-grained automatic evaluation benchmark trained across generation tasks, but more importantly I am curious about how these tools may offer new approaches to isolate sources of toxicity and bias in pre-training corpora and extend fine-grained RLHF [9] to broad behavioral alignment via multi-task fine-tuning with a diverse set of word-level evaluators. I am particularly excited in using span-based techniques to disentangle evaluation of logical and linguistic competence in open-source LLMs [3], which may offer a promising solution to peeling back imitation hidden within text generation [1].

Human Cognition as an Analogue for LM Behavior. While working on LLM analysis structured by linguistically motivated NLP evaluation, I developed an interest for a similar application of arguments in cognitive science motivated by our understanding of black-box reasoning and learning in human behavior as a complex system. E.g., when asked to describe recommendations for a medical diagnosis, do language models walk through their learned long-term memory like human experts, or is memory organization simply a reflection of entity occurrences in training data? In ongoing work with Prof. Sashank Varma in the Georgia Tech School of Psychology, I find evidence for the latter by drawing parallels between human experiments of memory organization and LM behavior through category fluency. Using data collected with human subjects, when asked to list animals as quickly as possible, their list typically contains a few similar examples, then a pause before jumping to a different subcategory. While we first show the same category grouping behavior can be observed qualitatively in lists generated by language models, we show the entropy of the next token probability distribution is directly tied to a jump in humans' average response time. Yet, models explore category structures in an exhaustive depth-first fashion rather than humans far more random walk through memory. We further challenge existing theories of human cognition which assume human category structure is essentially encoded as a non-contextual HMM to showing autoregressive LM predictions conditioned on a human-written sequence prior are far more representative of experimental data. With these observations, we argue that despite being better predictors of human category structure, language models during generation do not monitor long-term memory retrieval using a central executive similar to human behavior. We further incorporate this analysis to show the emergence of category structure and symbolic concepts as LM training progresses and evaluate the ability of de-biasing and de-toxicity interventions to reform category structure, which may have broad implication across memory retrieval tasks including long-form QA and dialogue systems. This work has led me to believe the rich human data and experimental designs in cognitive science may structure our higher-level claims of intelligence in LLMs spanning theory of mind and analogical reasoning by drawing comparisons to analogy and transfer in human cognition. Extending arguments showing LLMs encode biases of language (e.g., recursion) by pre-training on synthetic grammars [4], I am particularly interested in exploring whether instruction fine-tuning encodes biases of reasoning by adapting the study of progressive formalization in human learning. For instance, perhaps training on synthetic math and syllogism problems described in different stages of abstract (i.e., a defined context-free grammar) to concrete (i.e., natural language examples) may reveal the nature of abstraction in task generalization.

Language Modeling with a Central Executive. LLMs have a well-known inability to generalize to tasks requiring long-term planning or implicit reasoning, yet surprisingly, simply adding a deliberate mechanism for monitoring behavior (e.g., Chain-of-Thought) results in dramatic improvement [5, 8]. At AWS, I experienced this peculiar behavior first-hand when building an interactive account diagnostic tool: the LLM chat model only became helpful once I used a complex prompting setup to decouple information retrieval, symbolic problem solving and long-term planning using separate rule-based tools which removed any requirement for functional competence from the language model. These prompt-based approaches to reasoning are effective for solving isolated tasks, but require heavy human intervention and only patch a glaring hole in language models' lack of executive control. Informed by the two research directions discussed above, I hope to move beyond formulating planning and strategy as a brittle prompt-based search, and have begun work in this direction by using trained evaluators to guide generation. As part of an ongoing project in the NLP X lab at Georgia Tech, I show high quality constrained generation (e.g., machine translation, code generation) with LLMs can be achieved simply by decoding a candidate set of many LLM generations with a diverse prompt bank and using existing automatic neural metrics [2, 6] to select a target with Minimum Bayes Risk (MBR) decoding. With this additional layer of control, I show multi-prompt MBR can normalize LLM performance across any set of task descriptions or in-context examples, relaxing the generation requirement of the language model to estimating an adequate space of candidate generations by relying on a trained metric to discriminate quality sequences. The experience highlighted the reciprocal benefits between NLP analysis and development, and I plan to continue incorporating insights from the prior to build quantitative approaches to *control* behavior, which may offer a more trustworthy and cooperative interface for LLM applications. To this end, I am curious whether the same prompt-based executive control may also be constructed by limiting language model outputs to a pre-defined grammar which constrains formal reasoning behaviors within a task-specific search space and building value metrics trained to guide language generation with a learned task-independent abstract reasoning mechanism.

References

[1] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary LLMs. *arXiv:2305.15717*, 2023.

[2] Mounica Maddela, Yao Dou, **David Heineman**, and Wei Xu. LENS: A Learnable Evaluation Metric for Text Simplification. In *Proceedings of ACL*, 2023.

[3] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: A cognitive perspective. *arXiv:2301.06627*, 2023.

[4] Isabel Papadimitriou and Dan Jurafsky. Pretrain on just structure: Understanding linguistic inductive biases using transfer learning. *arXiv:2304.13060*, 2023.

[5] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of ACL*, 2023.

[6] **David Heineman**, Yao Dou, and Wei Xu. Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA. In *Proceedings of EMNLP*, 2023.

[7] **David Heineman**, Yao Dou, and Wei Xu. Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation. In *Proceedings of EMNLP: System Demonstrations*, 2023.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*, 2022.

[9] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Proceedings of NeurIPS*, 2023.