University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Computer Science

"**Controlled Event Generation in Large Language Models through Knowledge Localization, Meta-Reasoning, and Evaluator-Guided Reflection.**"

by **Bhiman Kumar Baghel**

Date: November 16, 2026
Time: 11:30 a.m. to 1:30 p.m.
Place: 5317 Sennott Square, 210 S Bouquet Street, Pittsburgh PA 15260

Committee:

- Xiang Lorraine Li, Assistant Professor, Computer Science, School of Computing and Information
- Adriana Kovashka, Associate Professor, Computer Science, School of Computing and Information
- Ryan Shi, Assistant Professor, Computer Science, School of Computing and Information
- Tessa Warren, Professor, Department of Psychology, Kenneth P. Dietrich School of Arts & Sciences

Abstract:

As language models increasingly inform everyday decisions, they must understand the nuances of real-world events. Some events are common and others rare, and rare events are not all alike: "a car repairing a child" is noise, since no line of reasoning makes it sensible in real life, whereas the reverse, "a child repairing a car," is uncommon but still possible. Yet these two examples are both rare in a model's training data, so frequency alone cannot separate noise from the uncommon yet possible. Generating and evaluating rare events therefore requires models to reason over world knowledge, not just recall frequent patterns. This thesis improves that reasoning in two complementary ways: verbalized reasoning, in which the model reasons explicitly in text, and internal reasoning, in which it reasons implicitly within its parameters, drawing on the knowledge stored there.

Psycholinguists have studied human event knowledge using events with graded plausibility. Inspired by this, we first define the task of controlled event generation and evaluation, in which a model must both generate and judge sets of such events. We find that even strong models rarely produce correctly graded event sets. Corrective feedback from an evaluator model improves their results, but its lessons are forgotten by the next example. To retain them, we distill many failures into reusable verbalized reasoning rules, which we first test on tool use, where objective outcomes make success easy to verify. Verbalized rules, however, are not fully effective, which we hypothesize is partly because the model's internal knowledge has not been updated to match them. Model editing can update a model's internal factual knowledge. We find that these edits are often imprecise. We developed method for making them more precise. Aligning verbalized and internal reasoning is essential for reliable deployment. Therefore, we propose to combine them in controlled event generation and evaluation. Unlike in tool use, here many events can fit the same plausibility grade, and judging them is subjective, so the signals for distilling reusable rules are noisy. We will adapt the rules to this setting and extend model editing so that the parameters hold the plausibility grade those rules state. Together, this work

aims to deepen language models' understanding of events, from common to rare, and make both their generations and evaluations more faithful to the plausibility grade of those events.