Detecting Factual Errors of Large Language Models

Language Models (LMs) acquire parametric knowledge from their training process, embedding it within their weights. The increasing scalability of LMs, however, poses significant challenges for understanding a model's inner workings and further for updating or correcting this embedded knowledge without the significant cost of retraining.

Moreover, when using these language models for knowledge-intensive language understanding tasks, LMs have to integrate relevant context, mitigating their inherent weaknesses, such as incomplete or outdated knowledge. Nevertheless, studies indicate that LMs often ignore the provided context as it can be in conflict with the pre-existing LM's memory learned during pre-training. Conflicting knowledge can also already be present in the LM's parameters, termed intra-memory conflict.

This underscores the importance of unveiling exactly what knowledge is stored and its association with specific model components, and how this knowledge is used for downstream tasks.

In this talk, I will present our research on evaluating the knowledge present in LMs, through a unified knowledge attribution framework, as well as diagnostic tests that can reveal knowledge conflicts.