Can an imitation learning agent inherit the goals of an expert: A case study of the traveling salesperson problem

Ognjen Stefanović, Trinity College

Language models (LMs) have been argued to possess beliefs, desires, and intentions which influence their outputs, resulting in arguments that these models can learn and pursue goals. This dissertation questions this narrative, asking whether LMs can learn to model goals and behave in a goal-directed manner when trained via imitation learning under an expert agent. Specifically, this case study focuses on the traveling salesperson problem (TSP), where the goal is unequivocally defined: to find the shortest route to visit each city and return to the start. The TSP setup used has five cities named `A`, `B`, `C`, `D` and `E`, which are placed on a 2D integer grid of shape 6×6. All possible city placements and their respective optimal paths are determined and used to create various training and evaluation sets. Each TSP example, along with its optimal traversal, is presented to the model as a sequence of characters and letters.

Under this clear and well-defined setup, the notion of goals and what it means for a LM to act in a goal-directed manner is explored. The definition of goal-directedness assumed here is that the model learns to perform a task and can exhibit *adaptivity*. Model adaptivity means the model will take "active" actions to complete the task even when there is a systematic difference between the train and test set. The experimental setups ensure that the goals in the evaluation sets are out-of-distribution in distinct ways while maintaining the correctness of the goals in the training set. This approach allows us to investigate generalisation in different problem setups.

The first experiment shows the model's ability to generalise to a previously unseen coordinate, achieving performance on par with the baseline. Additionally, it reveals that introducing this coordinate into the training set a few times does not affect performance.

Following this, tougher generalisation problems with more unseen coordinates at the board centre are posed. The model is evaluated on various problems containing cities at these previously unseen coordinates. The primary findings from these experiments indicate that the model acts in a goal-directed manner, successfully solving TSP with cities situated on previously unseen grid segments.

The goal-directedness of the model is also examined in presence of two different types of noise in the training sets. Random noise is introduced to the optimal salesperson traversing by randomly swapping two out of the last four cities in the optimal traversing. Systematic

noise is introduced by swapping cities `B` and `C` in the optimal salesperson traversing. Both noise types make the traversings suboptimal.

When random noise is introduced, the model demonstrates the ability to pursue the goal despite substantial noise. With systematic noise, the model tends to follow the goal most prevalent in the training dataset. *By doing so, the model is able not only to match but also to exceed the training agent's performance.* Additionally, we notice that the model generalises overconfidently on the evaluation set, finding optimal paths even when trained with significant amounts of systematic or random noise.

Finally, a mechanistic interpretability analysis of model activations in a simplified setting (for tractability purposes) with various inputs is conducted to try to uncover any emergent goal representations within the model weights. While the findings from this analysis are encouraging, there is still potential for future work.