Sommario

La Teoria della Mente (Theory of Mind, ToM) è essenziale nelle interazioni umane e ha implicazioni significative in vari campi. Questa tesi indaga il livello di ToM di cinque moderni Large Language Models (LLM) e propone un framework per misurare la complessità dei test legati a essa. Introduciamo il discorso con una breve panoramica storica, tracciando l'evoluzione della ToM dalla filosofia e psicologia fino alla sua influenza nella scienza informatica. Presentiamo in seguito una misura di complessità che distingue tra stati necessari e spurii, influenzanti la difficoltà di specifici problemi di ToM. Ispirandoci da questo framework, attingiamo idee dai Modelli del Mondo (World Models) per sviluppare Modelli del Mondo Discreti (Discrete World Models, DWM): descrizioni degli stati dell'ambiente create dai modelli di linguaggio stessi. Concludiamo con un'analisi dell'efficacia delle tecniche proposte, che migliorano le prestazioni fino al 9,99% nel miglior scenario possibile, e un ulteriore analisi sulla coerenza della misura teorica proposta con dataset pubblici. In conclusione, discutiamo le future direzioni per il benchmarking delle abilità di ToM negli LLM e forniamo un'analisi della memorizzazione dei dataset utilizzati.