research, such as mechanistic interpretability and human value/prefe

The project is motivated by a long-standing philosophical concern: can AI systems think, and how do we determine? Philosophers such as Hubert Dreyfus, John Searle, and Ned Block have provided critiques of early AI paradigms, particularly symbolic AI, by arguing that rule-based systems do not amount to real intelligence. As the paradigm shifts towards connectionism, represented by modern machine learning, these philosophical challenges have evolved.

My summer research focused on the intersection of philosophy, cognitive science, and artificial intelligence (AI), exploring how the current paradigm of AI relates to philosophical issues about mind, language, and knowledge. The primary goal was to conduct background research and identify a promising direction for my philosophy honors thesis.

The research utilized a combination of literature review and theoretical analysis. First, I conducted an in-depth examination of both classical and contemporary philosophical works relating to AI, ranging from Alan Turing's

In parallel, I reviewed the technical literature on AI, from the foundations of neural networks t
the architecture underlying LLMs. I specifically explored how these models operate, focusing
such as tokenization, embedding, backpropagation, the attention mechanism, and reinforceme
human feedback that enable these models to process and generate coherent language. In addit

alignment. I aimed to explain these concepts in accessible terms while maintaining philosophi
analyzing their significance.

general intelligence (AGI),