Externalized reasoning oversight is an alignment proposal that involves making an LLM externalize its reasoning (via chain of thought) so that its thoughts are visible to an external overseer, who then verifies that the LLM is not thinking deceptive thoughts.

This scheme requires:

- Not too much steganography: the model can't hide its reasoning in non-English logical thought. This may be the default of what would happen if we apply reinforcement learning (RL) to this model.
- A safe overseer that can look through the chain of thought reasoning and identify misaligned behavior.

A key dynamic is that natural human languages are not ideal to reason in; an AGI could be more capable by reasoning otherwise than in English. Applying RL to a model exerts pressure towards ideal logical reasoning, while self-supervised learning (SSL) exerts some pressure back towards just doing text prediction, and hence more of its reasoning being in human languages. A key question for this research agenda is thus how to keep the reasoning externalized, instead of collapsing into internal, more efficient/logical reasoning that is much harder to oversee.

Related

- What is Eliciting Latent Knowledge (ELK)?