arXiv.org

A General Language Assistant as a Laboratory for Alignment

Given the broad capabilities of large language models, it should be possible to work towards a general-purpose, text-based assistant that is aligned with human values, meaning that it is helpful, honest, and harmless. As an initial foray in this...