アラインメントチューニング

アライメントとは

- 言語モデルは人間の価値観や思考を考慮しないため、意図しない挙動を示すことがある
- 事前学習などとは全く異なる基準を考慮する必要がある
  - 役に立つ
  - 正直
  - 無害