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

人間によるフィードバックの収集

- 人間のラベラーの選択

  ➡ 一定の教育を受け、英語に 長けている適切な者を選ぶ

- 人間のフィードバックの収集

  ❖ ランキングベース：出力の候補に対してランク付けする

  ❖ 質問ベース：出力の有用さを測る質問に答える

  ❖ ルールベース：出力が一定のルールに適合するかテストする