ChatGPT

- 2022年11月にリリースされたGPTモデルをベースにした会話モデル
- 人間が生成した会話 (ユーザ役とAI役の両方を演じたもの) を使用してトレーニング
- 膨大な知識の蓄積・推論スキル・文脈の保持・人間の価値観への適合により、人間とのコミュニケーションで優れた能力を示した
- プラグインなどを組み合わせればさらに能力を拡張できる

GPT-4

- 2023年3月にリリースされた入力をマルチモーダルに拡張したモデル
- GPT-3.5よりも複雑なタスクを解く能力が高く、多くの評価タスクで大幅な性能向上を示している
- 6ヶ月間の反復調整により、悪意や挑発的な意図のある質問により安全に応答できるようになった
- Red Teamを組んでバイアスや憎悪に満ちた出力を生成させるなどの攻撃をさせ、有害なコンテンツ生成を減らした

A.Ovadya, “Red Teaming Improved GPT-4. Violet Teaming Goes Even Further,” *Wired*. Accessed: Aug. 20, 2023. [Online]. Available: https://www.wired.com/story/red-teaming-gpt-4-was-valuable-violet-teaming-will-make-it-better/