教育における生成AI活用推進リーダープログラム

生成AIの仕組み

機械学習

吉田塁

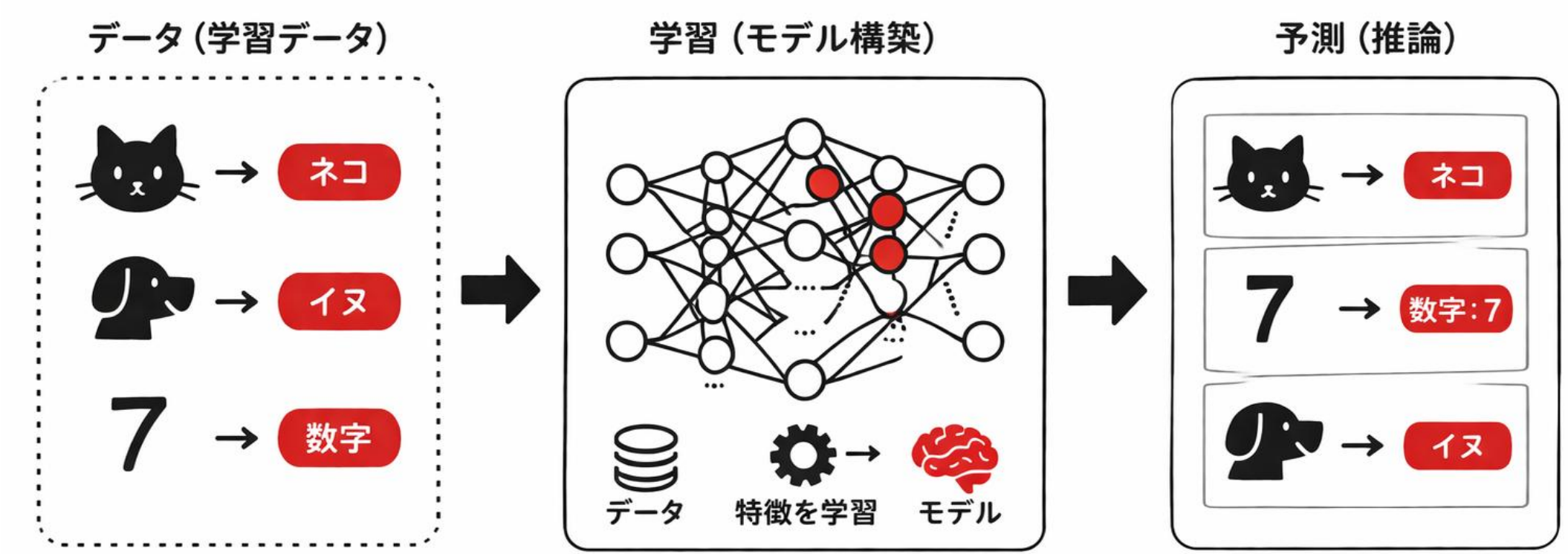
東京大学 大学院工学系研究科 准教授

LLM 寄附講座 特任准教授

2026年4月

機械学習

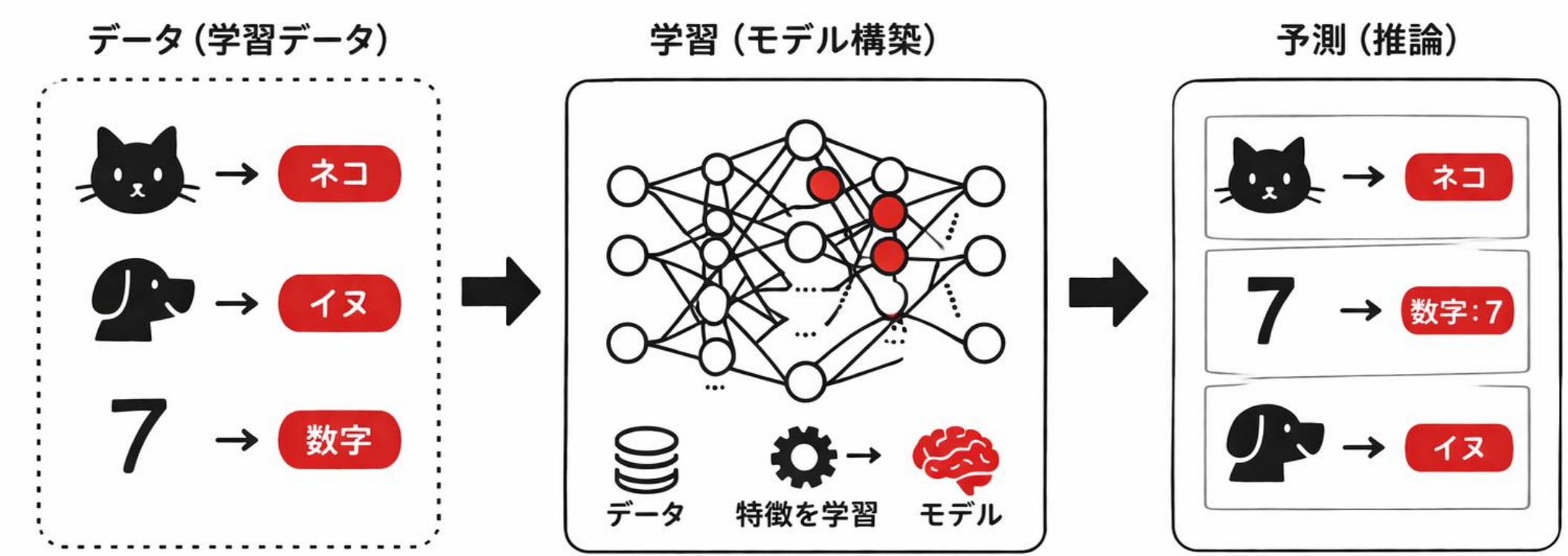
- **データから規則や傾向を学び、予測や判断を自動化する技術**
  - 教師あり機械学習、教師なし機械学習、強化学習など



画像は ChatGPT 5.4 Thinking を用いて作成

教師あり機械学習

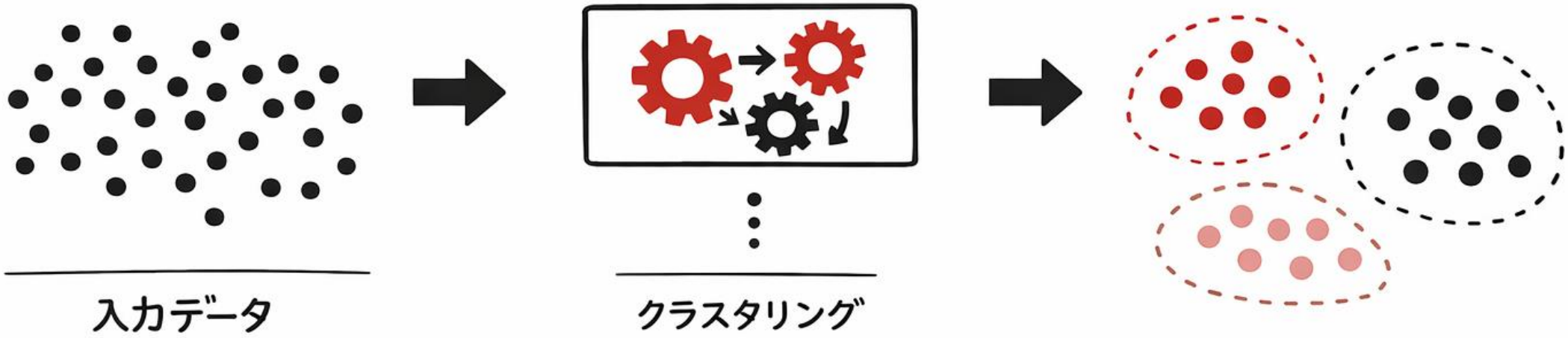
- 正解付きのデータを使って入力と出力の対応関係を学習し、新しいデータの予測や分類を行う手法



画像は ChatGPT 5.4 Thinking を用いて作成

教師なし機械学習

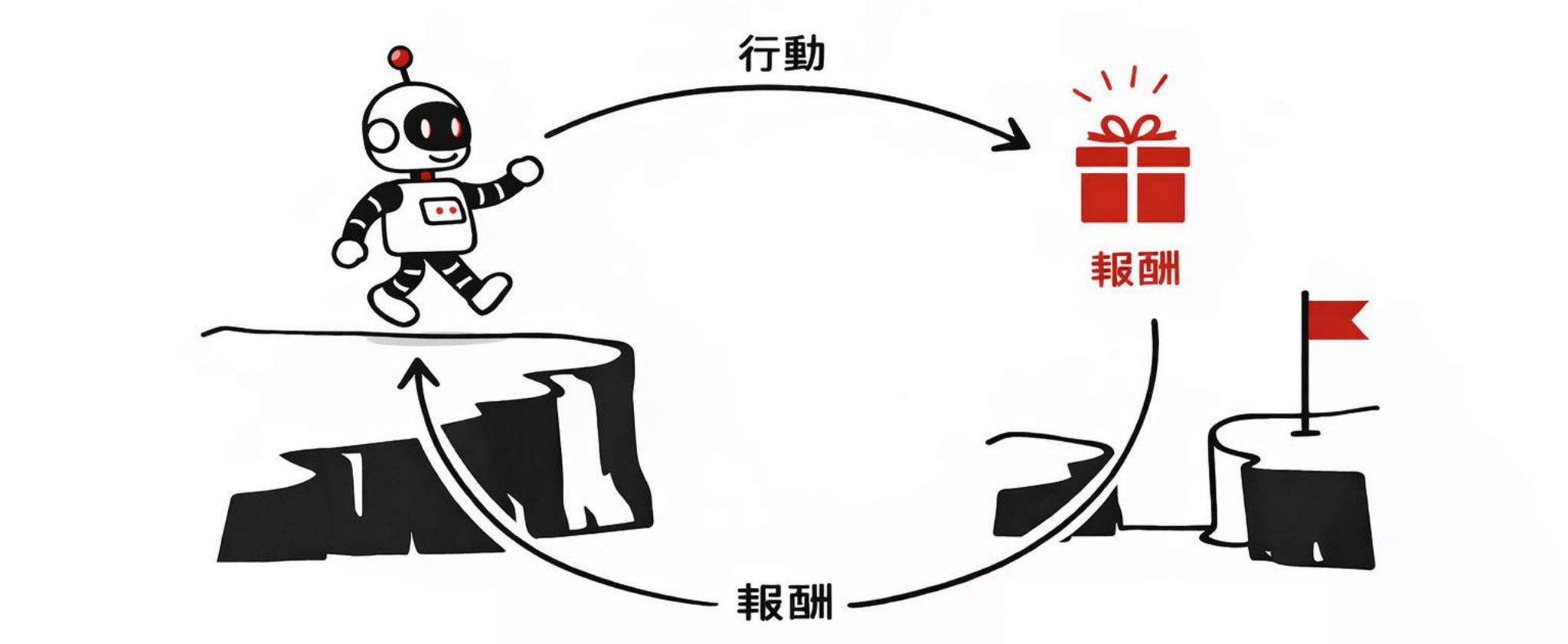
- **正解のないデータをもとに、似たもの同士のグループ分けや特徴的な構造を見つける手法**
  - クラスタリング、次元削減、異常検知など



画像は ChatGPT 5.4 Thinking を用いて作成

強化学習

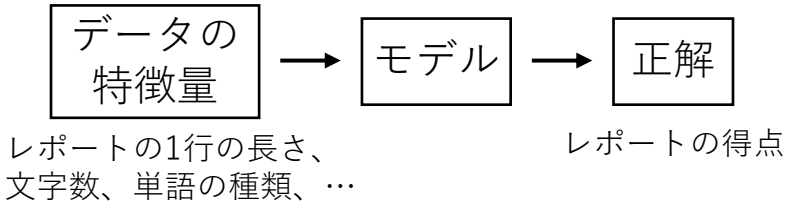
・行動の結果として得られる報酬をもとに、よりよい行動を学習していく手法



画像は ChatGPT 5.4 Thinking を用いて作成

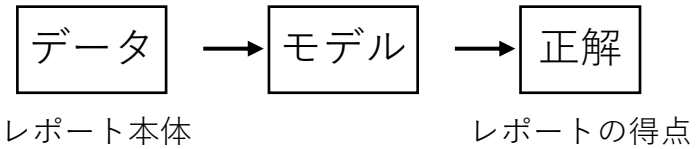
自然言語処理のパラダイム変化 (参考: Liu et al. 2023)

教師あり学習（非ニューラルネットワーク） (1950s-)



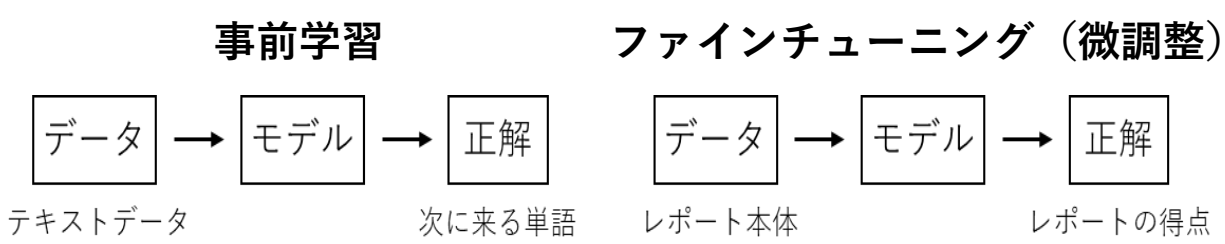
人間が特徴量（データに関する数値。例: レポートの長さ）を手作りして、モデルを学習させる

教師あり学習（ニューラルネットワーク） (2010s前半-)



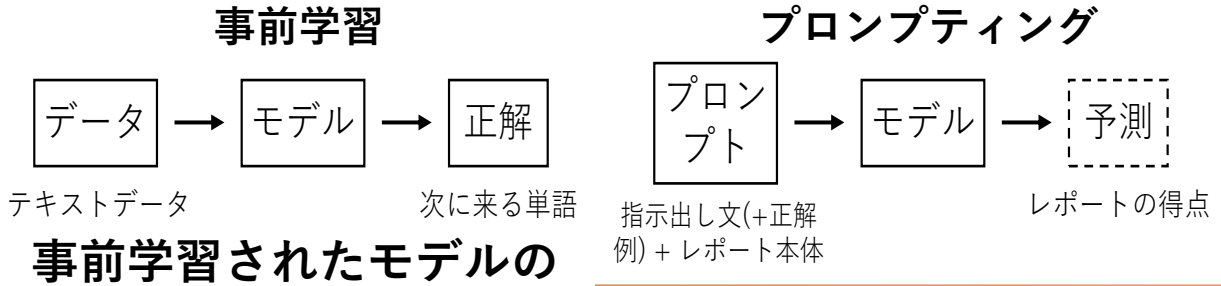
モデルが自動で特徴量を作ってくれる
必要なデータ数: 数百万等

事前学習とファインチューニング (2017/18-)
Transformer, BERT, GPT



事前学習されたモデルに、特定のタスクに特化させる微調整を行う
データ数: 数百〜数千等

事前学習とプロンプティング (2020-) GPT, PaLM

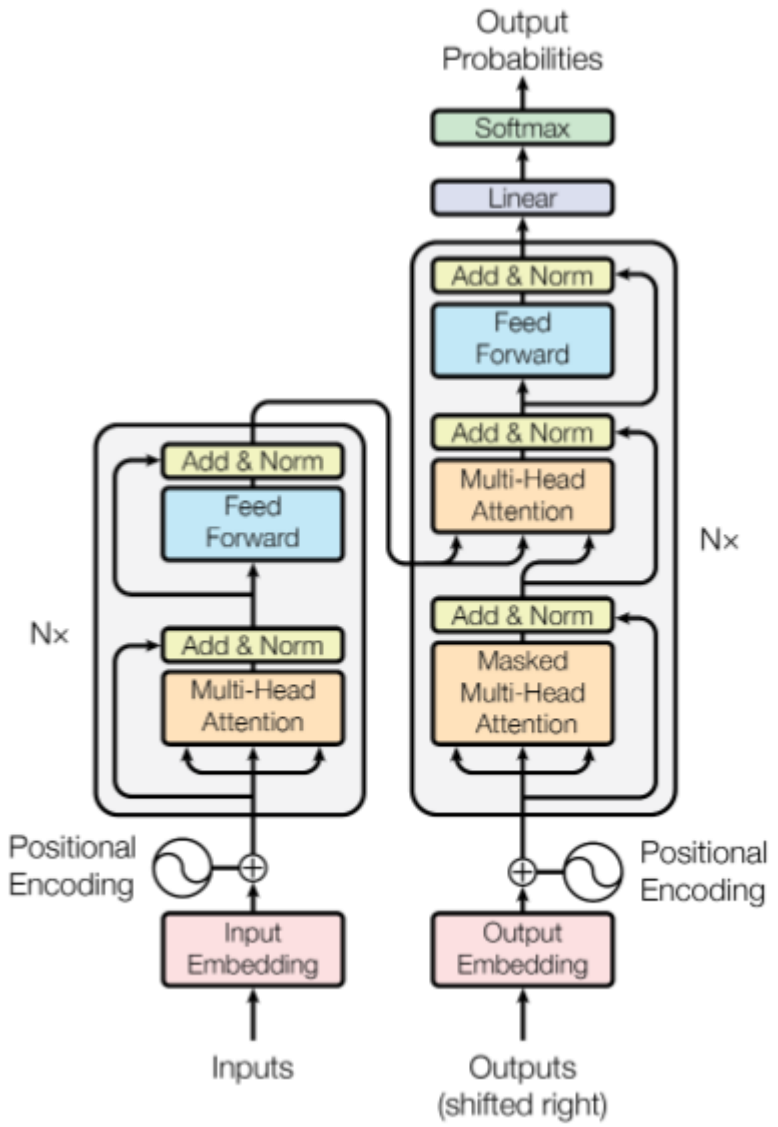


事前学習されたモデルに、特定のタスクに特化したプロンプトで操作
データ数: 0〜数個等

事前学習されたモデルの汎用性が高まっている

Transformer (Vaswani et al. 2017)

- **革命的なモデル**
  - 近年の主要な生成 AI の基盤となる
- **特徴（ざっくり）**
  - 従来はサブ的に使われていた Attention という機構にフォーカス
  - 離れた単語・文章の関係を把握可能
  - 学習計算の並列化が可能 (大規模なデータを処理可能に)



GPT-3 (GPT: Generative Pre-trained Transformer, Brown et al. 2020)

- **OpenAI によって開発された言語モデル（ざっくりいうと、入力されたテキストをもとに次に来る文字や単語の確率を計算するモデル）**

570GB (約3,000億単語) 分のテキスト

次に来る単語を予想して正解か判断

- **特徴**
  - 最大1,750億個とパラメータ数が多い
  - わずかな例示を含むプロンプトによって多様なタスクをこなす（いくつかタスクで当時の最高性能を叩き出す）
- **課題と対応策**
  - 人の意図に沿わない、信頼性の低い、攻撃的な出力を行ってしまう → 人の意図に沿った（人にアラインメントされた）学習を行おう（InstructGPT, ChatGPT）

ChatGPT 開発の流れ（ざっくり） https://openai.com/blog/chatgpt

Step 1

Collect demonstration data and train a supervised policy.

人好みの会話の出力になるように GPT-3.5 をファインチューニング

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

出力に対して「人の好み度」をわかるような報酬モデルを作成

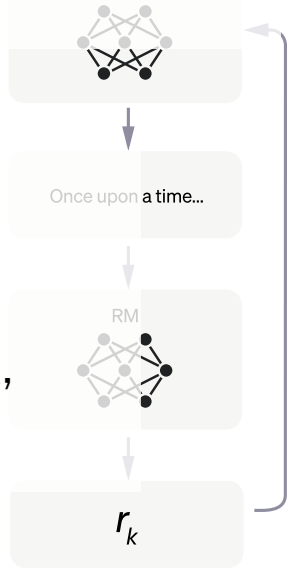
This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

ファインチューニングしたモデルを「人の好み度」報酬モデルを用いて強化学習

The PPO model is initialized from the supervised policy.



The policy generates an output.

The reward is used to update the policy using PPO.

（超ざっくり）
人が好む会話ができるように学習させたよ
(RLHF: Reinforcement Learning from Human Feedback, 人間のフィードバックによる強化学習)

参考文献

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.