AI エンジニアリング

基盤モデルを用いた AI アプリケーション開発の基盤と実践

ツールの使い方をメインのチュートリアルではない。ツールを選択するためのフレームワークを提供する。

第 1 章

AI エンジニアリングは、AI モデルを使ってアプリケーションを構築うるプロセス

言語モデルの出力は自由形式(オーペンエンド)、**補完マシン**と考えることができます。言語モデルの基本単位はトークンです。GPT-4 の場合、平均的なトークンは単語の長さの 3/4 です。

自己教師あり学習：モデルが学習するためのより大規模のデータセットの作業を可能にします。

OpenAI 初の生成系事前学習済み（Transformer）(GPT)モデルが 2018 年 6 月に発表された当初、そのパラメータ数は 1 億 1700 万個です。

大規模言語モデル（LLM）から基盤(Foundation)モデルへ
言語から図形への拡張だけではなく、Foundation 本来の意味合いが含まれている。

基盤モデルは、タスク特化（例えば、翻訳）型モデルから汎用モデルへの移行という側面も持っています。

データベースを使って指示（プロンプト）を補強する方法は、検束拡張生成（Retrieval-Augmented Generation：RAG）と呼びます。

プロンプトエンジニアリング、RAG、ファインチューニングは、モデルを特定のニーズに合わせて適応させるための、AI エンジニアリングにおける **3 つ**の主要なテクニックです。

リファクタリング（refactoring）とは、コンピュータプログラミングにおいて、プログラムの外部から見た動作を変えずにソースコードの内部構造を整理することである。

リポジトリとリポジトリ登録
リポジトリとは、データやファイルを一元的に保管・管理する場所を指す。
リポジトリ登録とは、大学や研究機関が作成した学術論文や研究報告などの電子データを、機関リポジトリと呼ばれるシステムに収集・保存し、インターネットを通じて公開することを指す

Moat（モート）堀。ビジネスの文脈では、競合他社の参入を防ぎ、自社の競争優位性を維持するための障壁を比喩的に表す。

レイテンシー（latency）とは、データ転送や処理において、要求を出してから応答が返ってくるまでの遅延時間を指す。

マイルストーン（英: milestone)は、プロジェクト管理において遅延が許されない節目となる工程のこと。

MMLU（Massive Multitask Language Understanding）

AI 推論　入力から出力を計算する時間

GDPR（General Data Protection Regulation：一般データ保護規則）とは、EU 域内の個人のデータ保護を目的とした法律です。EU データ保護指令よりも厳格で、EU 加盟国に直接効力を持ちます。

FOMO（Fear Of Missing Out、フォーモ、取り残されることへの恐れ）とは、「自分が居ない間に他人が有益な体験をしているかもしれない」、と言う不安に襲われることを指す言葉である。

（ビジネス）エコシステムとは、もともと生物学用語の「生態系」生態系、英:business ecosystem または digital ecosystem を指し、特定の領域で生物が相互に依存し共存する状態を意味します。ビジネス分野では、企業や製品、サービスなどが連携し、新たな価値や収益を生み出す仕組みを指します。

ファインチューニング(tine-turning 微調整)とは公開されている学習済のモデルに、独自のデータを追加で学習させ、新たな知識を蓄えたモデルを作り出す技術。追加学習とも呼ばれる。

キュレーション（Curation）**とは**、インターネット上に存在する膨大な情報を独自の基準で収集・選別・編集し、情報に新しい価値を付加した状態で共有することです。

アノテーション（Annotation）とは、AI に学ばせるための教師データを作る工程のことを指します。

第 2 章　基盤モデル

Gunasekar（人の名前）らは、70 億トークンの高品質なコーディングデータで、訓練した 13 億パラメータのモデル。Llama3、15 兆トークン（Llama2、2 兆トークン）。

ハルシネーション（hallucination)、または幻覚（げんかく）、でたらめ、作話（さくわ、confabulation)、ディルージョン（妄想、delusion）とは、人工知能によって生成された、虚偽または誤解を招く情報を事実かのように提示する応答のことである。

4C, Colossal Clean Crawled Corpus　（データセットに関連）

クリックベイト（Clickbait）とは、ネット上の虚偽・誇大広告の形態の一つで、ネットユーザーの興味を引くような文面のテキストやサムネイル画像を用いてリンクを踏ませ、欺瞞的な内容のコンテンツを読ませたり、見せたり、聞かせたりするものである

サムネイル（thumbnail、サムネールとも表記される）とは、主に画像や動画や文書などを表示する際に、一見でその内容を大まかに把握したり、複数を一覧表示する際に視覚的に素早く区別したりすることができるよう、縮小させた見本となる画像のこと。

Reddit（レディット）は、アメリカ合衆国発の掲示板型ソーシャルニュースサイトです。ユーザーはニュース記事、画像、テキストなどを投稿し、コメントや投票を通じて議論に参加できます。

RNN（Recurrent Neural Network：リカレントニューラルネットワーク）は内部に循環をもつニューラルネットワークの総称・クラスである。

エンコード(encode,符号化、暗号化)とデコード(decode、復元)

「プレフィル」(Prefill)とは、主にコンピュータやネットワークの分野で使われる用語で、データや情報を事前に自動で入力することを指します。

「Como」は主にスペイン語で使われる単語で、文脈によって様々な意味を持ちます。疑問詞としては「どのように、どんなふうに」という意味で使われます。
MLP（Multi-Layer Perceptron（感知)、

Transformer アーキテクチャ
アーキテクチャの変遷　seq2seq（Sequence to Sequence）=> Transformer（並列と重み）
アテンション（attention）は、認知的な注意を模倣するように設計された手法である。
Transformer アテンション機構
クエリベクトル（Q)、キーベクトル（K)、バリューベクトル（V)

FLOPS（フロップス、'Floating-point Operations Per Second）はコンピュータの性能指標の一つ

パラメータとハイパーパラメータ：ハイパーパラメータ（超母数、Hyperparameter）は、推論や予測の枠組みの中で決定されないパラメータのことを指す。

エポック数とは、機械学習において訓練データ全体をモデルに何回繰り返して学習させるかを示す回数です。

スクレイピング：これは「こする」「削り取る」を意味する英語の「scrape」に由来し、プログラムを用いてウェブページから必要な情報を効率的に取得することを指します。