生成AI基礎

8 生成AIを取り巻くツールと技術

GPTモデルの概要

- GPTとは：Generative Pre-trained Transformer
- Transformerアーキテクチャに基づく
- 人間のような自然な文章生成が可能

GPTモデルの種類

- **GPT-1**:
  - 2018年発表
  - 1億1700万のパラメータ
- **GPT-2**:
  - 2019年登場
  - 15億パラメータ

GPTモデルの種類

- **GPT-3**:
  - 2020年リリース
  - 1750億パラメータ
- **GPT-4**:
  - 2023年発表
  - マルチモーダルモデル（テキストと画像）
  - 詳細なパラメータ数は非公開

GPTの応用

- **ChatGPT**:
  - インタラクティブな対話型AIチャットボット
  - 人間のような自然な会話が可能
- **OpenAI API**:
  - 多様なタスクの自動化
  - チャットボット、コンテンツ生成、データ分析など多岐にわたるアプリケーションに応用

Claude 3の概要

- **Claude 3 Haiku**
  - 最速かつ最も手頃な価格
  - 迅速かつ正確な応答が必要なタスク向け
  - カスタマーサポート、大規模データ分析

Claude 3の概要

- **Claude 3 Sonnet**
  - バランスの取れたパフォーマンス
  - 信頼性が求められるエンタープライズの作業負荷向け
  - データ処理、セールスオートメーション、JSON生成
- **Claude 3 Opus**
  - 最も強力
  - 複雑な自動化、研究開発、戦略分析向け
  - 高度な推論、コーディング、多言語対応能力

Claude 3の応用

- **claude.ai**
  - Claude 3モデルを利用したウェブプラットフォーム
  - ユーザーフレンドリーなインターフェース
  - 多言語対応と強力なビジョン機能
- **Claude API**
  - 開発者向けのAPI提供
  - さまざまなアプリケーションに統合可能
  - 高速処理と大きなコンテキストウィンドウ

Geminiの概要

- **Geminiモデルの種類**
  - Gemini 1.0 Ultra
    - 57の科目で人間の専門家を上回る性能（MMLUベンチマーク）
  - Gemini 1.5 Pro
    - 性能と効率が強化
    - 1百万トークンまでの長文コンテキスト理解機能
    - テキスト、画像、音声、ビデオを含む複雑なマルチモーダルタスクに対応

Geminiの応用

- **Gemini Advanced**
  - Ultra 1.0モデルを使用
  - GmailやGoogleドキュメントの機能強化：メールの要約、ファイル検索
- **Gemini API**
  - 開発者向けのAPI提供
  - さまざまなアプリケーションに統合可能

プログラミング補助ツール

プログラミング補助ツール

- **GitHub Copilot**:
  - GitHubとOpenAIが共同開発したAI搭載のコード補完ツール
  - 機械学習モデルを使用して、リアルタイムでコード提案を提供
- **Amazon CodeWhisperer**:
  - Amazonが提供するリアルタイムのコード提案ツール
  - 機械学習を活用して、コメントや既存コードに基づいたコードを生成
- **Gemini Code Assist**:
  - Google Cloudが提供するAI搭載のコーディング支援ツール
  - コード生成や最適化、コンテキスト理解に優れる

GitHub Copilotの主な機能

- **コード提案と補完**:
  - コメントや部分的なコード入力に基づいてコードスニペットや関数全体を生成
  - ボイラープレートコードの自動生成でコーディングプロセスを高速化
- **AI搭載のチャット**:
  - 自然言語で質問し、コードの説明や提案を受けることが可能
- **Copilot Workspace**:
  - 計画、コーディング、コラボレーションのための包括的な環境
  - GitHub Codespacesと統合され、効率的な開発プロセスを支援

Amazon CodeWhisperererの主な機能

- **リアルタイムコード提案**:
  - コメントやコードを入力すると、リアルタイムでコードスニペットや関数を生成
- **セキュリティスキャン**:
  - 生成されたコードと手書きのコードをスキャンし、セキュリティ脆弱性を特定
  - OWASPトップ10などのベストプラクティスに準拠
- **個人利用が無料**:
  - すべての開発者が無料で利用可能
  - 無制限のコード提案、月50回のセキュリティスキャンをサポート

Gemini Code Assistの主な機能

- **コード生成**:
  - 大規模なコードベースの生成、最適化、翻訳をサポート
  - 複雑なタスクの自動化に対応
- **コンテキスト理解**:
  - Gemini 1.5 Proモデルは最大100万トークンの情報を処理可能
  - 大規模なコードベースを扱い、コンテキストに基づいた提案を提供
  - 新しい開発やレガシーコードの移行に非常に強力

大規模言語モデルを実行する基盤

Hugging Face

- 機械学習アプリケーションを作成するためのツールを開発
- モデルのホスティングとファインチューニング
- データセット共有プラットフォーム
- 効率的なモデル展開手法（例：量子化技術）

Amazon Bedrock

- **基礎モデル（FMs）**
  - Anthropic(Claude)、Cohere、Meta、のLLMsへのアクセス
  - 多様なアプリケーションに利用可能
- **モデル評価**
  - 異なるモデルの評価と比較が可能
  - 特定のユースケースに最適なモデル選択

ベクトルサーチエンジン

Elasticsearch

- 機械学習を利用して非構造化データを数値ベクトルに変換
- 近似最近傍（ANN）検索を使用して迅速な類似データ検索を実現
- ベクトル検索とキーワード検索を組み合わせたハイブリッド検索機能

Qdrant

- 高度なアルゴリズムで効率的かつ正確な近似最近傍検索を実現
- ベクトルに追加のメタデータ（ペイロード）を保存し、複雑なフィルタリングやカスタムビジネスロジックをサポート
- 水平スケーリングに対応し、大規模データセットや高負荷クエリにも効率的に対応

大規模言語モデルを拡張するためのツール

LangChain

- 大規模言語モデル（LLMs）を活用したアプリケーションの開発を支援する強力なフレームワーク
- モジュラーアーキテクチャ：感情分析、翻訳、名前付きエンティティ認識などの言語処理タスクのモジュールを追加およびカスタマイズ可能
- コンテキスト管理：高度なメモリ機構を通じて会話の一貫性と関連性を保持
- LangChain Expression Language (LCEL)：複雑なチェーンの作成を可能にする強力なツール
- 外部サービスとの統合：スピーチ・トゥ・テキスト、追加の機械学習モデル、画像認識やデータ分析などのクラウドベースのAPIと統合
- 高度なデータ取得とインジェスト：テキストスプリッターと埋め込みモデルをサポートし、大規模なデータセットの処理とクエリ処理を効率化

Dify

- 大規模言語モデル（LLM）を利用したアプリケーション開発のためのオープンソースプラットフォーム
- ワークフローオーケストレーション：AIワークフローを設計・テストするためのビジュアルキャンバス
- モデルサポート：GPTやLlamaなどの専有モデルおよびオープンソースモデルと統合
- プロンプトIDE：プロンプトの作成および比較のためのインターフェース
- RAGパイプライン：PDFやPPTなどの文書取り込みと検索のための強化された機能
- エージェントフレームワーク：LLM機能呼び出しやカスタムツールを使用したエージェントの定義。Google検索やWolframAlphaなど、50以上の組み込みツールを含む