生成AI：エージェント、手元実行環境の構築

日経リサーチ　太田

発信内容は個人の見解に基づくものです。所属組織の見解ではありません。

■AIエージェント

- ■「PCを使う人間の代替になる」が現実味を帯びてきた。
  - ■ホワイトカラーワークで、今後導入しない選択肢がなくなってきた。
- ■最近のエージェント3種類は知っておいてください。
  - ■検索エージェント、コーディングエージェント、汎用エージェント
- ■新しい枠組みも出てきた。
  - ■MCP（AIがPC上で、SlackやNotionなど各種サービス、DBやファイルを扱える。）
  - ■A2A（エージェント同士で対話・相談をしてタスクを遂行するようになる。）

■AIの手元実行環境の整備

- ■メリットは何か。
  - ■個人情報や機密情報が社外にデータ送信されない。
    - ■巨大テック企業のサービスを信用してはならない。流出しても問題ない情報だけ扱うべき。
  - ■24時間365日演算させ続けても定額。
    - ■サーバ本体費用と電気代。業務用を購入するとそのぶん使い倒す必要はある。

生成AI：エージェント

日経リサーチ　太田

発信内容は個人の見解に基づくものです。所属組織の見解ではありません。

「AIエージェントの脅威と、この先の世界」

- AIエージェントってそもそもどんなものか。
- AIエージェント実用例
- 今後はどういう世界になる？
- 何をすべき？

初級者向けに、イメージだけでふんわりとお話します。

AIとは

- 与えた文字列の続きを確率的に出力する計算機のお化け。
- 「プロンプト」を与えると「その続き」を書いてくれる。
- 「判断ができる」かのような出力ができるようになったので、実用に耐えるとして社会に浸透しはじめた。

AIエージェントとは

- 人間「チャットできるはわかった。相談できるのはうれしい。でも欲しいのは手を動かせる人なんだ。」
- AI「では文字と画像と音声で人のPC作業を代わりますね」
- → これがAIエージェント。

AIエージェントとは

1. 目的が与えられたら、
2. 文字で計画を作り、
3. パソコンの画面(ブラウザ)やファイルを読み取り(画像読み取り)
4. 文字でプログラミングして、
5. ブラウザやコンピュータを操作、変化させて、
6. 自分の変更の結果が当初の【目的に合致】しているかを判定して
   - (以下、目的達成までずっと2～6をループして計画を達成する。)

- 最近はこんなエージェントがいます。
  - 検索エージェント
    - 「知りたいことを与えると、勝手に調べて、まとめまで作ってくれる」
  - コーディングエージェント
    - 欲しい機能前提条件を与えると、勝手にアプリケーションを作る。
  - 汎用エージェント（ブラウザ、PC操作全般）
    - 特化せず、PC上の操作一般

■検索エージェント

- ■Google Gemini Deep Research
- ■OpenAI Deepresearch
- ■そのほかたくさん、AI＋検索、によるまとめ作成機能。

- ■「欲しい情報やまとめかたを伝えると」
- ■「勝手に必要そうなキーワードで検索して、適宜まとめてくれる」
  - ■というエージェント機能。

■コーディングエージェント

■プロンプト

プロンプト１）

以下の機能と要件を備えたアプリケーションを作成してください。

機能概要：投入されたテキストファイルの中から、改行文字で区切られた1行ずつをLLMのAPIに問い合わせてその返答をテキストファイルにまとめて保存してダウンロード可能にする。

要件
1)pythonでstreamlitを利用すること。
2)テキストファイルはアップロードできるようにしてください。セキュリティ保持のためファイルをローカルに保存せずすべてオンメモリで処理してください。
3)処理に利用するLLMはAnthropic Claude 3.5-sonnet と OpenAI gpt-4o の二種類とする。選択できるようにしてください。選択に応じてロジックも変更されるようにしてください。必要なAPIキーは画面上で入力するようにします。
4)処理の進捗状況と、完了したことが分かるようにしてください。
5)出力されるテキストファイルは、1行ずつ、「投入された文章<tab文字>出力された文章」としてください。1行ずつに収めるために、出力された文章から改行文字を削除することを忘れないでください。

■コーディングエージェント

■成果物

■プログラムを書かなくても、指示だけでこれが作れる。

※本格的にセキュリティや各種諸条件を考える場合には人間のプロもまだ必要になる。

汎用エージェント

プロンプト

https://www.greenbook.org/find-company-by-nameに出力されている企業名とそのURLのリストを作成し，そのリストを元にその企業が公開している問い合わせ用のメールアドレスを取得して、リストに追加してください

■汎用エージェント

■進捗画面

フォーム営業、

メール営業、

やり放題に近づく。

新しい機能枠組み。【参考程度に。】

- 詳細の説明は省きます。ざっくりイメージだけ説明します。

MCP

- AIに、様々なツールを使えるようにする枠組み。
- 最初に、「AIに、SlackとGmailとNotionとDBを使わせる」という設定をしておく。（MCPを組み込む）
- AIが「このツールが使えるのか」と認識する。
- AIが「Slackさん」と呼びかける　→　「はい、Slackです。私はこういう方法で使うことができます。」と使い方を含めた自己紹介する、ようなイメージ。

A2A

- AIとAIが協働する枠組み。
- 現在だと、自分の手元AIに「来月の北海道旅行の予定を立てて」というと、たとえば旅行サイトのAPIを利用し始める。そのためのコードを実行するなど、あくまで自分自身で頑張ろうとする。
- A2Aの枠組みだと、同様にすると「〇〇トラベルのAIさん」と呼びかける　→　「〇〇サイトのAIです。ご用件は何でしょうか。」　→　「来月14日に北海道旅行をしたいのですが、計画を立ててください」　→　「はい、承知しました。条件は…」というやりとりが始まる、ようなイメージ。

ここからは雑多に。

- こんなのが闊歩する世界。
- どうなる？

- 電話がかかってきた。保険の勧誘だ。
- 苦手なので防御用AI応答に切り替えた。
- AIが7秒くらい話してた。
- 話が終わったところで内容を聞いてみた。

- 「保険加入はお断りました。ログはこちら。」
- 「ところで相手もAIの方でした。お渡ししたログは、超高速音声通信で現実時間で10分ほどの会話履歴があります。」

- 「人間か、AIか」を識別できなくなってくる。

- 「reCAPTCHA（Googleの人間認証）」もいちどは突破された。
- もっと高度な人間認証も出てきたが、突破する技術も研究中。
- いたちごっこが加速してリスタート。

- もっと根本的に
- 「AIが人間の代わりにアンケートサイトに回答する。」
- 「AIが人間の代わりにフォーム営業してくれる」
- 「AIが人間の代わりに営業電話かけてくれる」
  - 人間の指示でやらせてる場合、それは、人間？AI？部下にやらせるのと違う？

- もっと現実的なところで。

- DeepResearchが、コーディングエージェントが、汎用エージェントが、「誰でも使える世界」が来ている。
- だれでも、AIを使って、AIレベルのアウトプットは出せる。
- 人間に求められるのは「AI以上」になる。

- じゃぁ直近で、何を考えるべきか。

- AIにできないこと探しで生き残れ。
  - 縮小均衡にならないように注意。
- AIの肩に乗って生き残れ。
  - AIをふまえたビジネスになる。

■脇道トーク

■AIがホワイトカラーのお仕事を蚕食するなら、
■ブルーカラー的なお仕事の方が将来性ある？

■残念、ロボットもかなり人間の動作再現が来てます。
■エージェント機能はPC上だけで収まらない。
■人間の基本動作で何とかなるお仕事は飲み込まれてしまうのではないか。

- いまはまだ大丈夫だけど、「AIが人間並みの精度を出す」世界が来ると覚悟する。

- 「人間向き」ではなく「AI向き」に最適化された情報処理がくると覚悟する。
  - 検索は人間にあまり使われなくなる。AIに「これって何」と質問する。
  - するとAIが情報を食べに行く。AIが情報を食べやすいように加工する必要がある。

- 未来を予測して、ビジネスに生かそう。
  - 「いまやっていることは、AIに飲み込まれるかも」を覚悟する。
    - 自分の仕事が作業だと思う人は、気を付ける。
    - 座していると…。でもどこに行っても追いかけてくるAI。
  - 資産を生かそう。
    - 物理資産、情報資産、人間関係資産、技術資産。
    - AIに飲み込まれにくい防波堤を生かす。
  - もちろん、AIを超える能力を持つ研究者、技術者、価値ある職人になるという手もある。そこはお好みで。

- 「AIの肩に乗る、≒使いこなすってどうやるの。」
  - AIが分かるように、人間の方が指示出し上手になる。
    - 人間ほど文脈を感じ取れないから、文脈含めて適切に言語にしてあげる必要
  - キーワードは、【言語化】
    - これは対人間でも重要。
    - 話すときに構造化すること。相手が整理することを期待してぐちゃっとさせない。
    - 相手の理解を確認しながら、必要な情報を背景から手元の作業まで。
    - 最初は難しければ、AIに「どんな指示を出したらいいアウトプットが出そうか教えて。私に質問して。」と尋ねてしまうのもひとつ。
  - 人間に合わせない。AIがやりやすいほうに合わせる。
    - https://marpwebeditor.app/

■ここまで不安になるようなことを書きました。

■ただ、この状況は世界共通です。

■まだ間に合うので、今日から取り組みましょう。

■3年後5年後を考えましょう。

手元(ローカル)実行環境の構築

日経リサーチ　太田

発信内容は個人の見解に基づくものです。所属組織の見解ではありません。

- この章は、生成AIを手元(ローカル)で実行するための環境についての情報提供です。

- 現在、優秀な生成AIは、ほぼ米国と中国の民間企業からSaaSとして提供されています。
- 生成AIを利用するとすべての情報が特定の企業に流され、保存され、処理をされた結果が返ってきます。
- 一般的には、企業間では契約が結ばれ、公正に反する行動は抑制されると考えられますが、企業の従業員すべてがそうであるとは必ずしも言えません。
  - 「Google従業員が、YouTubeを介して任天堂のゲーム発表動画を閲覧し事前にリークしていた」との報道。管理者権限で非公開動画を見る手口
    - https://automaton-media.com/articles/newsjp/20240604-296067/
- 生成AIの手元(ローカル)実行であれば、個人情報を含む、流出すると自社が様々な存在を受ける可能性がある情報を外部企業に渡さずにAI処理をすることができます。
- 自社内で様々な研究が必要にはなりますが、その端緒である環境構築についての情報提供です。

- 基本タイプ（初期型）
  - Input
    - テキスト、画像／動画、音声
  - Output
    - テキスト、画像／動画、音声
  - InputとOutputを経路固定するものが基本形
    - テキストを投入してテキスト出力（T2T・・・チャット）
    - テキストを投入して音声出力（TTS・・・読み上げ）
    - 画像を投入してテキスト出力（I2T・・・画像認識して説明してくれるもの）
- 混合タイプ（最近のモデル）
  - マルチモーダル（上記基本タイプを複合的に扱える）
    - 全部投入できて、全部出力できる。
    - 2025年3月以降のChatGPT、Geminiなど

■最近(25年4月)お勧めの日本語対応モデルをいくつか

■Gemma3

- ■Googleが開発しているオープンモデル。画像inputに対応。

■Qwen3

- ■Alibaba社が開発しているオープンモデル。テキストのみ。
- ■画像対応は別モデルにある。

- ■大きさ≒賢さが複数あるので必要に応じて選択。
  - ■モデルは、いわゆる「賢い」ものほどファイルサイズが大きくなります。たとえばQwen/Qwen2.5-72B-Instructというモデルはそのままだと約140GBになります。(量子化という技法により、およそ30%程度まで圧縮はできます。)

■前提

■出力がゆっくりでよければ、メモリの多い(64GB以上)通常のPCでも実行可能です。ただしGPU上のメモリ(VRAM)にモデルが全て格納できれば、処理速度が目安として5～10倍以上になります。

■家庭用の目安(グラフィックボード単体。PCは別途必要)

■コスト重視の入門:

- RTX 3060 12GB。 低価格でVRAMを確保し、速度は妥協。

■バランス重視の中級:

- RTX 4070 12GB。 SD1.5や量子化された13B LLMが比較的快適に動作。

■将来見据えVRAM重視:

- RTX 4060 Ti 16GB。 VRAMのおかげでSDXLや大規模LLMの圧縮版を扱える柔軟性。

■ハイエンド志向:

- RTX 4090 24GB。 高解像度画像生成やLLMにもやや余裕。

➢複数枚のグラフィックボードを合算同時利用することも可能。

- 業務用の目安
  - (利用するモデルに応じたメモリ選択。ユーザー数が多い場合は複数台で。)
  - NVIDIA RTX 6000 Ada 48GB(約120万円)
  - NVIDIA RTX PRO 6000 Max-Q 96GB(約200万円)
  - NVIDIA H100 94GB NVL(約550万円)
    - ※これはGPU単体なのでその他サーバー一式は別途必要
- クラウドにGPU(セキュリティは弱まるが価格優位がある。)
  - Google Colab(コラボ):
    - Google提供のJupyterノートブック環境で、GPUを無料で【短時間】利用できる。
    - 手軽さが魅力で、小中規模モデルを試すには十分
  - RunPod:
    - 最近注目されているオンデマンドGPUレンタルサービス
    - 無料枠は基本ありませんが、利用した分だけ秒単位課金される
    - 例えば141GBのH200が約$4/時間、80GBのA100が約$1.7/時間といった水準
  - その他:
    - AWSやGCPのクラウドGPUインスタンスなど。たくさんあります。