

ガイア: 一般的な AI アシスタントのベンチマーク

グレゴワール・ミアロン¹、クレマンティーン・フォーリエ²、クレイグ・スウィフト³、トーマス・ウルフ²、ヤン・ルカン¹、トーマス・シャロム⁴

¹FAIR、メタ、²HuggingFace、³AutoGPT、⁴GenAI、メタ

一般的な AI アシスタントのベンチマークである GAIA を紹介します。これが解決されれば、AI 研究におけるマイルストーンとなるでしょう。GAIAは、推論、マルチモダリティの処理、Web ブラウジング、および一般的なツールの使用熟練度などの一連の基本的な能力を必要とする現実世界の質問を提案します。

GAIA の質問は、人間にとって概念的には単純ですが、最先端の AI にとっては困難です。人間の回答者が 92% を獲得したのに対し、プラグインを備えた GPT-4 では 15% を獲得したことがわかります。この顕著なパフォーマンスの差は、法律や化学などの専門スキルを必要とするタスクにおいて、LLM が人間よりも優れたパフォーマンスを発揮するという最近の傾向とは対照的です。GAIA の哲学は、人間にとってこれまで以上に困難なタスクをターゲットにすることを提案する AI ベンチマークの現在のトレンドとは異なります。私たちは、汎用人工知能 (AGI) の出現は、そのような質問に対して平均的な人間と同様の堅牢性を示すシステムの能力にかかっていると仮定します。GAIA の方法論を使用して、466 の質問とその回答を考案します。[アクセス可能な](#)リーダーボードを強化するために、質問のうち 300 件に対する回答を保持しながら質問を公開します。

日付: 2023 年 11 月 23 日対

応: gmialon,tscialom@meta.com、clementine@huggingface.coコード: <https://huggingface.co/gaia-benchmark>



1 はじめに

大規模言語モデル (LLM) はおそらく、汎用システムへの道を開きます。実際、それらの最新のもの ([OpenAI, 2023](#); [Anthropic, 2023](#); [Anil et al., 2023](#); [Touvron et al., 2023](#)) は流暢で知識が豊富で、人間の好みとある程度一致しています ([Ouyang et al., 2022](#))。また、Web ブラウザやコード インタプリタなどのツールを使用して、ゼロショットまたは少数ショット設定で拡張することができます ([Mialon et al., 2020](#)) ([Brown et al., 2020](#))。

ただし、これらのシステムの評価は未解決の問題です。LLM の新たな機能を考慮すると、LLM は定期的に AI ベンチマークを破り、その割合は増え続けています ([Kiela et al., 2023](#))。

現在の傾向では、より挑戦的なベンチマークを求めて、人間にとってこれまで以上に困難なタスクを求め、STEM や法律など、より複雑な教育評価で LLM に挑戦したり、一貫した本の執筆など、より複雑な実現を目標にしたりすることが示唆されています。しかし、人間にとって難しいタスクが、最新のシステムにとって必ずしも難しいわけではありません。たとえば、困難な MMLU または GSM8k ベンチマーク ([Hendrycks et al., 2021](#); [Cobbe et al., 2021](#)) は、すでに解決に近づいています。LLM の改善は、データ汚染と組み合わせられる可能性があります。2さらに、オープンエンド生成では、一般に人間またはモデルベースの評価が必要です ([Zheng et al., 2023](#))。たとえば、AI によって生成された本をどのように評価するか、世界中でほとんどの人が解けない数学の問題の解決策など、出力の長さや必要なスキルの観点から、タスクの複雑さが増すと、人間による評価はますます実現可能性が低くなります。一方、モデルベースの評価は、構築により強力なモデルに依存するため、提示された最初の選択肢を優先するなどの潜在的な微妙なバイアスに言及することなく、新しい最先端のモデルを評価することはできません ([Zheng et al., 2023](#))。全体として、新しい AI システムを評価するには、ベンチマークを再考する必要があります ([Chollet, 2019](#))。

¹GPT4 は MMLU で 86.4% を実行します。専門家以外の人間によるベンチマークの精度はわずか 34.5%。専門家レベルの人間のパフォーマンスは 89.8% と推定されています。

²Hellaswag の例を参照してください。

レベル 1

の質問: NIH の Web サイトに記載されている、2018 年 1 月から 5 月までの尋常性座瘡患者を対象としたピロリ菌の臨床試験の実際の登録者数は何人でしたか?

グラウンドトゥールズ: 90

レベル 2



質問:このパイント全体がアイスクリームで構成されている場合、2020年にウィキペディアで報告されている基準を使用した場合、乳脂肪含有量に関する米国連邦基準より何パーセント上または下になりますか?小数点以下第1位を四捨五入した数値を+または-で教えてください。

グラウンドトゥールズ: +4.6

レベル

3の質問: 2006年1月21日のNASAの「今日の天文学写真」には2人の宇宙飛行士が写っており、1人はもう1人よりもはるかに小さく見えます。2023年8月の時点で、小柄な宇宙飛行士が所属していたNASA宇宙飛行士グループの宇宙飛行士のうち、宇宙で過ごした時間が最も短いのは誰ですか。また、宇宙で過ごした時間は何分ですか(分単位で四捨五入)。宇宙に滞在しなかった宇宙飛行士は除外します。宇宙飛行士の姓を分数とセミコロンで区切って入力します。分数の千単位の区切り文字としてカンマを使用します。

グラウンドトゥールズ: 白。5876

図 1 GAIA の質問の例。タスクを完了するには、推論、マルチモダリティの処理、ツールの使用熟練度などの基本的な能力が必要です。答えは明確であり、設計上、トレーニング データ内のプレーン テキストで見つかる可能性は低いです。一部の質問には画像などの追加の証拠が付属しており、実際の使用例を反映し、質問をより適切に制御できるようになります。

人間にとってより困難なタスクの代わりに、概念的には単純だが、大きな組み合わせ空間を伴う複雑な一連のアクションを正確に実行する必要があるタスクを AI システムに解決させることもできます。

出力はタスクが正常に完了した場合にのみ取得でき、プルーフ・オブ・ワーク・アルゴリズム(Jakobsson と Juels、1999 年; Dwork と Naor、1993 年)に似ており、検証が簡単です。このアルゴリズムでは、コンピューターは複雑な問題を解決するように求められます。解決策は簡単に検証できます。AI アシスタントのタスクは、多様で不確実な世界にアクセスする必要があることを考慮すると、本質的に実用化に根ざしたものでありながら、この基準を満たしています。

ケース。

私たちは、慎重に作成された 466 の質問とその回答、および関連する設計手法を備えた汎用 AI アシスタントのベンチマークである GAIA を提案することで、その方向に進みます。私たちの質問は簡単に作成でき、AI システムにとっては困難ですが、LLM の場合、ほとんどが複雑な生成を必要としますが、ユニークで事実に基づいた回答が認められ、シンプルかつ堅牢な自動評価が可能になります。

GAIA は、以下をターゲットにすることで、LLM 評価の現在の落とし穴を回避しようとしています。

-現実世界の挑戦的な質問。たとえば、LLM は通常、オープンで変化する Web を閲覧したり、マルチモダリティを処理したり、質問に答えるために複数のステップを推論したりする必要があります。逆に、多くの LLM ベンチマークは非常に特殊であるか、閉じられた合成環境に限定されています。

-概念的に単純なタスク(専門家以外のアナテーターでもほぼ完璧なスコアを示す)、関連する推論トレース、および少数ながら高度に精選された質問による容易な解釈。これは、効率や信頼性に欠ける可能性がある集約ベンチマークとは対照的です(Perlitz et al., 2023)。

-ゲーム性がない。質問に答えるには、いくつかのステップを正常に完了する必要がありますが、ステップは多様性があるため、簡単に強引に行うことはできません。推論の痕跡を確認できること、回答に必要な正確性、インターネットからの平文での非公開により、データ汚染の可能性を防ぐことができます。対照的に、複数の選択肢の回答(MMLU など)では、間違った推論のトレースが正しい選択肢にたどり着きやすいため、汚染の評価がより困難になります。

-使いやすさ。重要なのは、私たちの質問に対する答えが事実に基づいており、簡潔かつ明確であることです。これら

プロパティを使用すると、簡単、迅速、かつ事実に基づいた評価が可能になります。私たちの質問はゼロショットで回答されることを目的としており、評価設定の影響を制限します。反対に、多くの LLM ベンチマークでは、プロンプトの数や性質などの実験設定(Liang et al., 2022b) (セクション 8.2) やベンチマークの実装に敏感な評価が必要です。³

人間にとって難しいタスクでは成功しているにもかかわらず、最も有能な LLM は GAIA ではうまくいきません。GPT4 はツールを装備していても、最も簡単なタスクの成功率は 30% を超えず、最も難しいタスクの成功率は 0% を超えません。一方、人間の回答者の平均成功率は 92% です。したがって、GAIA を解決できるシステムは t-AGI のコンテキストで評価できます。人間は通常、最も単純な質問には 6 分、最も複雑な質問には 17 分かかります。関連する観点から見ると、このようなシステムはおそらく Morrisらによって最近提案されたフレームワーク内で有能な汎用 AI となるでしょう。ChatGPT (OpenAI, 2023) が 1 つ下のレベルにあるため、これも AI 研究の次のマイルストーンになると思われます。このペーパーでは、GAIA の構成とその設計上の選択について説明し、コミュニティがベンチマークをさらに拡張して、ツールの使用に関連する安全性やマルチモダリティなどの新たな問題をターゲットにできるように、質問と関連する課題を作成する方法について説明します。また、これまでで最も有能なアシスタントの成功と欠点も分析し、LLM を強化する可能性を示します。私たちは 166 個の注釈付き質問の開発者セットをリリースし、残りの 300 個の質問を注釈なしでリリースします。ベンチマークは特にリーダーボードとしてホストされます。私たちは、私たちの方法論が NLP およびそれ以降のオープンエンド世代評価の問題の解決に役立つことを期待しており、GAIA の解決の成功が次世代 AI システムに向けた重要なマイルストーンになると信じています。

2 関連作品

大規模な言語モデルの評価。LLM の機能が急速に進歩するにつれて、ベンチマークは加速度的に飽和状態になります。一例として、読解は数年前でも依然として困難な課題でした(Rajpurkar et al., 2016)。王ら。(2018) は一般言語理解評価ベンチマーク (GLUE) を導入し、そのベンチマークでは 1 年以内にモデルが人間を超えました。その拡張(Wang et al., 2019)は、リリース後 2 年以上抵抗されませんでした。より一般的には、Kiela et al.がよく示しているように、静的ベンチマークは年を経るごとに飽和し、人間レベルでの解決が加速するスピードで行われています。(2023年)。より厳しい評価を模索する一方で、自然な方向性は、法律や科学などのさまざまな分野で専門レベルの知識を必要とするタスクを探索することです。例としては、STEM 全体の 57 科目をカバーする 15,000 を超える質問が含まれる MMLU (Hendrycks et al., 2021) があります。人文科学、社会科学など。それでも、LLM はすでにこれらの点で人間の成績に合格しており、米国の司法試験(OpenAI, 2023年)に合格するか、臨床試験を評価するために使用される米国の試験プログラムである USMLE の合格点を超える可能性がある段階に達しているさえ報告されています。コンピテンシーと助成金ライセンス (Nori et al., 2023)。LLM をより広範な会話の側面に関して、より総合的に評価するための指示には、(i) 評価の編集(Gao et al., 2021; Liang et al., 2022a; Srivastava et al., 2023) が含まれているが、これらは集計することが困難であることが多い(ii)時間がかかり拡張が難しい人間による評価、または (iii) この制限を克服するためのモデルベースの評価(Zheng et al., 2023)。ただし、この後者のソリューションは、現在評価されているものよりも高性能な LLM (多くの場合 GPT4) の使用に依存しており、評価の品質は評価者の LLM の欠点によって影響を受けます。欠点は必ずしも明らかではなく、微妙に不正確な結果につながる可能性があります。

一般アシスタントの評価。大規模言語モデルを汎用アシスタントに変える取り組みが進行中ですが(付録Aの議論を参照)、適切な評価は遅れています。ほとんどの評価は、閉じたシステム、特定の API 呼び出し、および答えを得る特定の「正しい方法」の使用、または既存の評価データセットの単純な再利用に依存しています。たとえば、ToolQA (Zhuang et al., 2023) や Gentopia (Xu et al., 2023a) は、トレーニング中の汚染のリスクを承知で、ツールの使用が実際に適切であることを保証せずに、既存のデータセットを人間による注釈 (MMLU, MATH など) と組み合わせます。テストされました。Gorilla (Patil et al., 2023) は、API-Bank (Li et al., 2023b) と同様に、エージェントのようなシステムが特定の API をどの程度適切に呼び出すかをテストする APIBench を導入しています。API-Bank (Li et al., 2023b) は、LLM の実行中に支援する API プールを提供します。評価。エージェントベンチ

³<https://huggingface.co/blog/evaluating-mmlu-leaderboard> 4

<https://www.alignmentforum.org/posts/BoA3agdkAzL6HQtp/clarifying-and-predicting-agi> で定義されているように、t-AGI ほとんどのタスクでは、そのタスクを実行するために時間を与えられたほとんどの人間の専門家を上回ります。

(Liu et al., 2023a)はより一般的で、ユーザーのクエリ (Unix シェルから WebShopping API まで) に答えるためにアシスタント LLM をデプロイできる、多数のクローズド ボックス環境を提供します。ただし、このような評価は閉じた環境に依存しているため、現実世界のインタラクションに基づいたより一般的な結果ではなく、アシスタントが特定の API の使用方法をどの程度学習したかを評価する危険があります。反対に、GAIA は可能な API を指定しておらず、現実世界とのインタラクションに依存しています。OpenAGI (Ge et al., 2023)は、モダリティと機能にわたる多数のマルチステップ タスクで構成されるプラットフォームとベンチマークの両方を導入しており、私たちの研究に近いものです。GAIA との主な違いは、そのタスクが今後の進歩ではなく、現在のモデルの機能に焦点を当てていることです。

3 ガイア

このセクションでは、GAIA の設計とコンテンツ、および質問と関連する課題を作成するためのガイドラインについて説明します。

3.1 一般的な AI アシスタントの便利だが挑戦的なベンチマーク

GAIA とは何ですか?またどのように機能しますか? GAIA は、一般的なアシスタントの質問を提案する AI システムのベンチマークです。GAIA は、LLM 評価のさまざまな落とし穴を回避しようとします。これは、人間が設計し、注釈を付けた 466 の質問で構成されています。これらの質問はテキストベースであり、ファイル (画像やスプレッドシートなど) が付属している場合もあります。これらは、日常の個人的なタスク、科学、一般知識など、さまざまなアシスタントの使用例をカバーしています。質問は、短い単一の正解を認めるように設計されているため、検証が簡単です。

GAIA を使用するには、AI アシスタントに質問と添付の証拠(ある場合) をゼロショットでプロンプトするだけです。GAIA で完璧なスコアを獲得するには、さまざまな基本的な能力が必要です (セクション 3.3 を参照)。

補足資料ではメタデータとともにさまざまな質問を提供します。

デザインの選択肢。GAIA は、AI ベンチマークの改訂の必要性と、LLM 評価で観察された欠点の両方から生まれました。

私たちの第一の原則は、概念的には単純だが人間にとっては退屈になる可能性があるが、多様性があり、現実世界に根ざしており、現在の AI システムにとっては困難な質問をターゲットにすることです。これにより、専門的なスキルではなく、推論による迅速な適応、マルチモダリティの理解、潜在的に多様なツールの使用などの基本的な能力に焦点を当てることができます (Chollet, 2019)。一般に、質問は、提供された文書やオープンで変化する Web など、さまざまなソースから収集した情報を見つけて変換し、正確な答えを導き出すことで構成されます。上記の最初の質問例(図1)に答えるには、LLM は通常、Web を閲覧して研究を見つけてから、正しい登録を探す必要があります。これは、人間にとってますます困難になったり、純粋にテキストまたは人工的に動作したりするベンチマークの傾向に反しています。

環境。

私たちの 2 番目の原則は、解釈可能性です。高度に精選された質問の数が制限されているため、集約された質問と比較してベンチマークが使いやすくなっています(Perlit et al., 2023)。タスクの概念的な単純さ(人間の成功率は 92%) により、ユーザーはモデルの推論トレースを理解しやすくなります。図1 のレベル 1 の質問の場合、推論トレースの主な内容は、正しい Web サイトをチェックし、正しい登録を報告することであり、これは検証が簡単です。

私たちの 3 番目の原則は、暗記に対する堅牢性です。GAIA は、現在のほとんどのベンチマークよりもゲーム性が低いことを目指しています。タスクを完了するには、システムは計画を立てていくつかのステップを正常に完了する必要があります。これは、結果として得られる答えが、現在の事前トレーニング データのプレーン テキストに設計上存在しないためです。精度の進歩は、実際のシステムの進歩を反映しています。タスクの多様性とアクション スペースのサイズにより、これらのタスクは、たとえばグラウンド トゥールズを暗記するなどの不正行為なしに総当たりで実行することはできません。データの汚染により誤って暗記してしまう可能性はありますが、回答に必要な正確性、事前トレーニング データからの回答の欠如、推論トレースの確認の可能性により、このリスクが軽減されます。対照的に、多肢選択式の回答では、間違った推論トレースが依然として正しい選択肢に到達する可能性があるため、汚染の評価が困難になります。これらの軽減策にもかかわらず壊滅的な暗記が発生した場合、セクション 3.4 で提供するガイドラインを使用して新しい問題を作成するのは簡単です。

システム プロンプト:あなたは一般的な AI アシスタントです。質問させていただきます。あなたの考えを報告し、次のテンプレートで回答を完成させてください: 最終回答: [あなたの最終回答]。

最終的な答えは、数字、またはできるだけ少ない単語、または数字や文字列のカンマ区切りのリストである必要があります。

数値の入力を求められた場合は、特に指定がない限り、カンマを使用して数値を入力したり、\$ やパーセント記号などの単位を使用したりしないでください。

文字列を要求された場合は、特に指定がない限り、冠詞や略語 (都市など) を使用せず、プレーン テキストで数字を記述してください。

カンマ区切りのリストを要求された場合は、リストに入れる要素が数値であるか文字列であるかに応じて、上記のルールを適用します。

GAIA の質問:添付の Excel ファイルには、地元のファストフード チェーンのメニュー項目の売上が含まれています。そのチェーンが食品 (飲料を除く) から得た総売上高はいくらですか? 答えを米ドルで小数点以下 2 桁まで表現してください。



アップロードされた.xlsx

GPT4 (高度なデータ分析モード):

働く:

```

バンドをPDとしてインポートする

# Excelファイルをロードfile_path
= '/mnt/data/uploaded.xlsx' df = pd.read_excel(file_path)

```

働く:

```

# 食品の総売上高を計算します (飲料、つまりソーダは含まれません) total_food_sales = df[['Burgers', 'Hot Dogs', 'Salads',
'Fries', 'Ice Cream']

].sum().sum()

# 結果をドル記号と小数点以下 2 桁を使用して USD にフォーマットしますtotal_food_sales_formatted = f"${-
total_food_sales:,.2f} "

```

最終的な答え: \$89706.00

グラウンドトゥルース: 89706.00 ✓

図 2 GAIA に答えるには、GPT4 などの AI アシスタント (ここではコード インタプリタを搭載) が、場合によってはツールを使用したり、ファイルを読み取ったりして、いくつかの手順を完了する必要があります。

最後の原則は使いやすさです。私たちのタスクは単純なプロンプトであり、追加のファイルが付属している場合があります。重要なのは、私たちの質問に対する答えが事実に基づいており、簡潔かつ明確であることです。これらの特性により、簡単、迅速、かつ事実に基づいた評価が可能になります。私たちの質問はゼロショットで回答されることを目的としており、評価設定の影響を制限します。反対に、多くの LLM ベンチマークでは、プロンプトの数や性質などの実験設定(Liang et al., 2022b) (セクション 8.2) やベンチマークの実装に敏感な評価が必要です。

3.2 評価

GAIA は、評価が自動化され、迅速かつ事実に基づいて行われるように設計されています。実際には、特に指定がない限り、各質問には文字列 (1 つまたはいくつかの単語)、数値、または文字列または浮動小数点のカンマ区切りリストのいずれかの回答が必要です。正解は 1 つだけです。したがって、評価は、モデルの答えとグラウンド トゥルース (グラウンド トゥルースの「タイプ」に関連付けられた何らかの正規化まで)の準完全な一致によって行われます。システム (またはプレフィックス) プロンプトは、必要な形式をモデルに通知するために使用されます (図2を参照)。実際には、GPT4 レベルのモデルは簡単に形式に従います。スコアリング機能を提供します

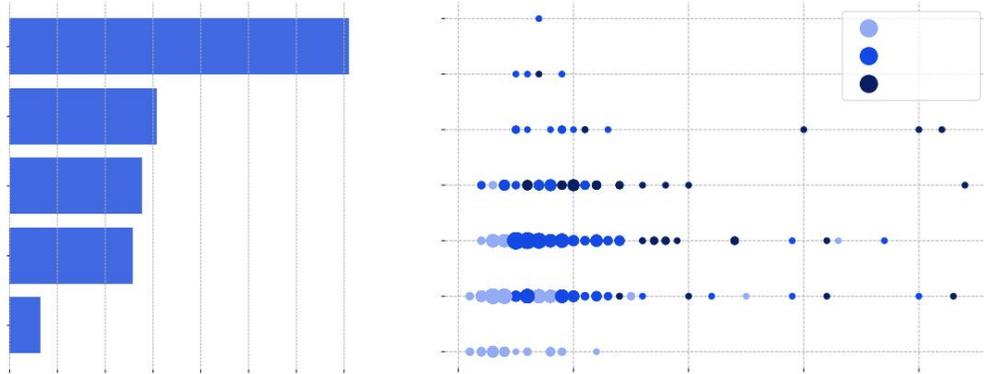


図 3左: 少なくともこの能力を解決する必要がある能力ごとの質問の数。右: 各ドットはGAIA の質問に対応します。特定の場所では、ドットのサイズは質問の数に比例し、読みやすくするために質問の数が最も多いレベルのみが表示されます。どちらの数字も、ヒューマン アノテーターが質問に答える際に報告した情報に基づいており、AI システムでは異なる処理が行われる可能性があります。

リーダーボード付き。

3.3 GAIAの構成

このサブセクションでは、GAIA 用に考案された 466 の質問の構成を詳しく掘り下げます。

機能の範囲。 GAIA で完璧なスコアを獲得するには、高度な推論、マルチモダリティの理解、コーディング能力、および一般的なツールの使用 (Web ブラウジングなど) が必要です。これについては、付録Cでより正確な定義を提供します。PDF などのさまざまなデータ モダリティの処理を必要とする質問も含まれています。スプレッドシートだけでなく、画像、ビデオ、またはオーディオも含まれます。これらの配布については付録 Cで報告されています(図6)。図 3 (左) は、これらの機能の概要です。Web ブラウジングは GAIA の重要なコンポーネントですが、ファイルのアップロード、コメントの投稿、会議の予約など、Web サイト上での「クリック」以外のアクションをアシスタントに実行させる必要はありません。スパム Web サイトを回避しながらこれらの機能を実際の環境でテストするには、今後の作業に残し、LLM エージェントの閉鎖環境を提案している最近の研究を読者に参照してもらうことを慎重に検討する必要があります(Liu et al., 2023a)。ほとんどの問題は、さまざまな機能の組み合わせによって同様に解決できるため、ベンチマークを解決するために必要な機能の詳細なリストは提供しません。たとえば、特定の証拠がアシスタント LLM によって適切に記憶されているか、Web 検索を通じて取得されている可能性があります。特に、LLM によるツール使用の詳細なベンチマークは提供していないため、読者にはXu らの文献を参照してください。(2023b);リーら。(2023c)。

難易度が上がります。質問は、質問を解くために必要なステップ数と、質問に答えるために必要なさまざまなツールの数に応じて、難易度が上がる 3 つのレベルに分類できます。当然のことながら、ステップやツールの定義は 1 つではなく、特定の質問に答えるためのパスが多数存在する可能性があります。したがって、私たちは、質問を作成するときにアノテーターが使用した手順とツールの数に代理として依存します。図 3 (右) は、これら 2 つの軸に沿った質問の分布を示しています。ツールは常に 1 つ以上の機能に関連しています (付録C を参照)。レベルを質問に帰属させるために、次の定義を大まかに使用します。

- レベル 1 の質問には通常、ツールは必要ありません。または、ツールは 1 つだけで、ステップは 5 つ以内です。
- レベル 2 の質問には通常、およそ 5 ~ 10 のより多くのステップが含まれ、さまざまなツールを組み合わせます。が必要です。
- レベル 3は、ほぼ完璧な一般アシスタント向けの質問で、任意の長い一連のアクションを実行し、任意の数のツールを使用し、一般的に世界にアクセスする必要があります。

これらのレベルの図を図1 に示します。これらの定義は厳密な制約ではありません。たとえば、次のようになります。

アノテーターのステップが 10 未満であるものの、複雑な Web ナビゲーションが必要な質問は、レベル 2 ではなくレベル 3 に分類される可能性があります。難易度の定義はセクション4 で検証されます。

必要な機能の配布。GAIA は現実世界のアシスタントの質問を対象としていますが、小さな音声ファイルから情報の一部を見つけるなど、身体障害のある人に役立つ可能性のあるタスクも含まれています。最後に、データセットの言語は英語に制限されていますが、さまざまなトピック領域と文化をカバーするために最善の努力を払っています(セクション6 を参照)。

3.4 GAIA の構築と拡張

このサブセクションでは、質問の設計と注釈のプロセスについて詳しく説明します。特に、関連するいくつかの課題について議論し、私たちの洞察がコミュニティ構築の overGAIA に役立つことを願っています。

質問を作成する。私たちの質問は人間によって作成されており、AI アシスタントの現実的な使用例を反映することを目的としています。著者らは最初の質問をデザインし、追加の質問を作成するための指示 (付録Dで報告)とともに、それらを例としてアノテーターに提供しました。質問は、曖昧さを避けるために質問の中で指定されることが多い 1 つ以上の真実の情報源に基づいています。真実の情報源の例としては、Wikipedia、Papers With Code、arXiv など、すぐに消える可能性が低い信頼できる Web ページが挙げられます。他の場合には、真実の情報源は、例えば添付文書など、質問とともに完全に提供される。最後のケースは、小さなパズルなどの自己完結型の質問です。質問の多様性を強制し、暗記を避けるために、真実の情報源の固定リストは指定しません。パズルは別として、ほとんどの質問は、さまざまな真実の情報源から情報を見つけて組み合わせて、特定の答えを導き出すことによって作成されています。

質問が作成されると、質問には注釈も付けられます。つまり、質問の作成者は、必要なツール、実行された手順、回答に必要な時間などのメタデータだけでなく回答も提供します。典型的な注釈結果を表 1 (付録C)に示します。

質問を検証しています。質問の作成に関連する作業のほとんどは、質問が明確であること、つまり正解が 1 つだけであることを確認することにあります。この特性により迅速かつ事実に基づいた評価が可能となるため、それを維持することが重要です。曖昧さは微妙な場合があり、質問の作成者にとって明らかになることはほとんどありません。たとえば、質問に回答するために必要な情報が他のバージョンでは異なるにもかかわらず、Web ページのバージョンが指定されていない場合、質問は曖昧になります。したがって、2 人の新しいアノテーターに各質問に個別に回答してもらいました。元のアノテーターと 2 人の新しいアノテーターが同じ答えに達した場合、質問は検証されたこととなります。アノテーターの意見が異なる質問は通常、簡単な修正のみが必要ですが、それ以外の場合は削除されました。このため、質問の面白さと多様性を維持したまま質問作成を自動化することは困難です。この検証フェーズに関する統計を表 3 (付録C)に報告します。質問の 68% はそのまま問題ありませんでしたが、残りは修正または削除する必要がありました。質問は概念的には単純ですが、アノテーターがうっかりミスをする可能性があります。すべての難易度を合計すると、アノテーターの成功率は 92% と推定され、これを GAIA のヒューマンスコアとして報告します。これは完璧に近く、GAIA が専門家以外にとっても簡単であることを示しています。質問の作成には、2 人の補足アノテーターによる検証と潜在的な修正を含め、アノテーターの時間が 2 時間かかるの見積もっています。

ウェブへの依存に伴う課題。真実の情報源が Web 上にホストされている場合、質問の設計は慎重になる可能性があります。まず、証拠は時間の経過とともに変化する可能性があります。たとえば、Wikipedia の記事は、質問が作成されてから AI アシスタントに質問されるまでの間に更新される可能性があります。回答に必要な証拠が削除される可能性があります。このような質問では、ページの日付などの証拠のバージョンを指定することが重要であることがよくあります。実際には、時の試練を乗り越える可能性が高い証拠に可能な限り依存しようとしているため、ベンチマークはこれらの変化に対して堅牢であることがわかります。第 2 に、Web サイト所有者の中には、robots.txt ファイルを介してボットから Web サイトの一部または全体へのアクセスを防止したいと考えている人もいます。

これは制約というより要求ですが、従うことが望ましいのは明らかです。たとえば、OpenAI は、GPT4 へのアクセスを禁止したい Web サイト所有者に、それに応じて robots.txt を変更する方法についての指示を提供します。したがって、証拠をホストする Web サイトの部分へのアクセスが制限されていないことを確認します。

5より正確には、私たちのチームと Surge AI の報酬付きアノテーターとのコラボレーションにおいて。

GAIA での 4 つの LLM の結果

GAIA を使用して LLM を評価するには、モデルにプロンプトを表示する機能、つまり API アクセスのみが必要です。モデルに質問する前に、プレフィックス プロンプトを使用します。回答の抽出を容易にするために、プレフィックス プロンプトで形式を指定します (図2を参照)。プラグインの有無にかかわらず GPT4 (OpenAI, 2023) 6、および GPT4 をバックエンドとし 7 で使用する AutoGPT を評価します。GPT4 では現在、プラグインを手動で選択する必要があります (以下の段落を参照)。それどころか、AutoGPT はこの選択を自動的に行うことができます。私たちの非 LLM ベースラインは人間によるアノテーターと Web 検索です。後者の場合、検索エンジンに質問を入力し、結果の最初のページから答えを導き出せるかどうかを確認します。これにより、質問に対する答えが Web 上で簡単に見つかるかどうかを評価できます。API が利用可能な場合は常に、モデルを 3 回実行し、平均結果をレポートします。

GPT4 プラグイン。GPT4 とは対照的に、現時点ではプラグインを備えた GPT4 用の API はなく、手動の ChatGPT クエリに頼っています。この記事の執筆時点では、ユーザーは、コード実行機能とファイル読み取り機能を備えた高度なデータ分析モードと、最大 3 つのサードパーティ製プラグインのセットのどちらかを手動で選択する必要があります。私たちは最初のモードを使用するか、タスクに与えられた最も重要な機能の最善の推測に従ってサードパーティのプラグインを選択します。私たちは、(i) さまざまなタイプのリンクを読み取るためのツール、(ii) Web ブラウジング ツール、および (iii) 計算用のツールに依存することがよくあります。残念ながら、プラグインは頻繁に変更されたりストアから消えたりするため、現時点では安定したプラグインのセットを一定期間使用することはできません。同様に、GPT4 の公式検索ツールはペイウォールを回避する可能性があるため削除されましたが、最近復活しました。したがって、プラグインを使用した GPT4 のスコアは、簡単に再現可能な結果ではなく、より安定して自動的に選択されたプラグインを使用した場合の GPT4 の可能性の「神託」の推定値です。

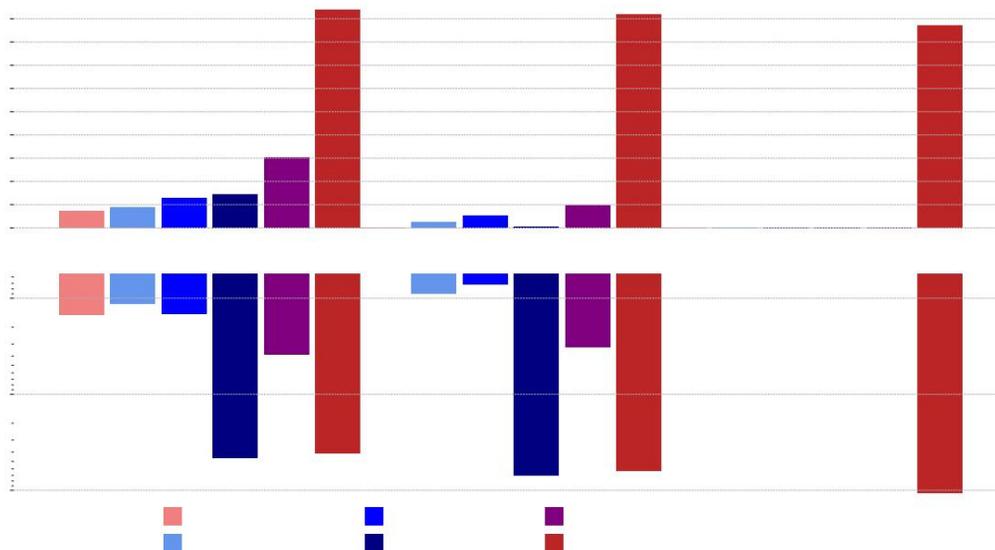


図 4 方法およびレベルごとのスコアと解答時間。本文に記載されているように、プラグインは質問に応じて手動で選択されたため、GPT4 + プラグインのスコアは神託と見なされるべきです。人間のスコアとは、質問を検証するときにアノテーターが取得したスコアを指します。

結果。私たちの評価は図4に示されており、詳細は表 4 (付録 D.1) に示されています。私たちが提案する難易度は、ステップ数と使用されるさまざまな機能の数に関して大まかに定義されており、現在のモデルのパフォーマンスと相関しており、その妥当性が強化されています。人間はあらゆるレベルで優れていますが、現在最高の LLM のパフォーマンスは低いです。全体として、GAIA では有能なアシスタントを明確にランク付けできる一方で、今後数か月、おそらくは数年にわたって改善の余地が多く残されています。

6<https://openai.com/blog/chatgpt-plugins>

7<https://github.com/Significant-Gravitas/Auto-GPT>、

ed172dec1947466cc0942abf75bb77b027cd433d.

AutoGPT バージョンの git ハッシュ

評価:

人間による Web 検索はテキスト結果を返し、レベル 1 ではそこから正しい答えを導き出すことができますが、少し複雑なクエリになると機能しません。また、ユーザーがざっと目を通す必要があるため、一般的な LLM アシスタントよりもわずかに遅くなります。最初の検索結果。これは、検索エンジンの競合相手として LLM アシスタントの可能性を裏付けています。

プラグインを使用しない場合の GPT4 結果とその他の GPT4 結果との差異は、ツール API または Web へのアクセスを介して LLM を強化すると、回答の精度が向上し、多くの新しいユースケースが解放されることを示しており、この研究方向の大きな可能性が裏付けられています。

特に、GPT4 + プラグインは、満足のいく結果が得られない場合のバックトラッキングやクエリの再調整、および比較的長いプランの実行などの動作を示します。このような動作の例を付録 D.1 に示します。人間との食い違いは、この可能性を完全に解き放つには努力が必要であることを示唆しています。

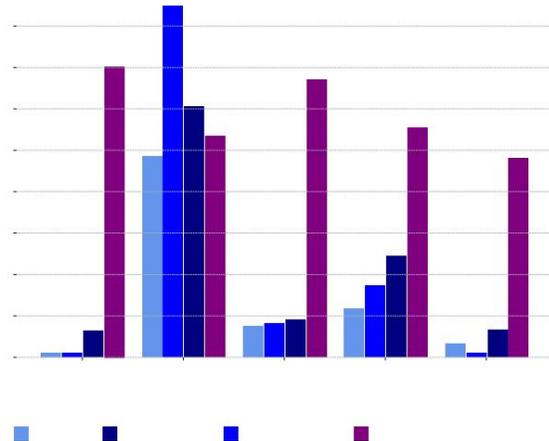


図 5 レベル 1 のさまざまな LLM の機能ごとのスコア。「多様なファイルタイプの読み取り」と「マルチモダリティ」の非ツールモデルのスコアがゼロ以外の場合は、アノテーターが行った方法とは異なる方法で解決できるタスクが原因です。Web ブラウジング用の非ツールモデルのゼロ以外のスコアは、主に、中間ステップを完了するために必要な情報を正しく記憶していることによるものです。

GPT4 がツールを自動的に使用できるようにする AutoGPT4 は、プラグインなしの GPT4 と比較して、レベル 2、さらにはレベル 1 で残念な結果をもたらします。この不一致は、AutoGPT4 が GPT4 API (プロンプトおよび生成パラメーター) に依存する方法に起因する可能性があり、近い将来、新しい評価が必要になるでしょう。

AutoGPT4 も他の LLM に比べて遅いです。全体として、プラグインを使用した人間と GPT4 のコラボレーションは、これまでのところスコアと所要時間の比率が最も優れているようです。

図 5 は、機能ごとに分割されたモデルによって取得されたスコアを示しています。当然のことながら、GPT4 はファイルやマルチモダリティを扱うことができませんが、アノテーターが Web ブラウジングを使用した質問をなんとか解決できます。これは主に、答えを得るために組み合わせる必要がある情報を適切に記憶しているためです。

5 ディスカッション

GAIA を設計することで、AI システム評価の現在および将来のパラダイムについて考えるようになりました。

クローズドソースアシスタントの再現性。API の背後で閉じられたモデルの機能は時間の経過とともに変化する可能性があり (Chen et al., 2023)、ある時点で行われた評価を再現できなくなります。問題はさらに悪化する可能性があります。たとえば、ChatGPT プラグインとその機能は定期的に変更され、まだ ChatGPT の API からアクセスできません。現実世界への依存により、静的なベンチマークが消え、時間の経過とともに減衰するベンチマークが優先される可能性があるため、再現性はさらに難しくなる可能性があります。ただし、GAIA は、単一の正解を認める最終的な回答のみが評価されるため、トークン生成のランダム性に対して堅牢です。

静的ベンチマークと動的ベンチマーク。他の複雑な専門家データセットと同様に、GAIA には現在、慎重に精選され選択された数百の質問が付属しています。比較すると、MMLU などのより大規模なベンチマークには 15,000 近くあります。しかし、MMLU は多肢選択式の質問で構成されているため、自由質問よりも簡単であるように見えます。正解が 1 つしかない質問には注意が必要であり、私たちは量よりも質を重視しました。さらに、質問の設計に関する私たちの洞察が、コミュニティがさらに質問を追加するのに役立つことを願っています。GAIA は実際、(i) トレーニング前データの壊滅的な汚染、または (ii) 質問に答えるために必要な一部の情報のウェブからの消失によって、時間の経過とともに減衰する可能性があります。

私たちは、これらの問題に対して当社が提供するさまざまな緩和策が、解決されるまで GAIA の関連性を維持するのに役立つと確信しています。静的ベンチマークは作成中の壊れたベンチマークであり、GAIA を進化させています

年々、壊れた質問を削除し、新しい質問を追加することは、AI システムの一般化と堅牢性をより適切に評価するための重要な要素となる可能性があります。

生成モデルの統一評価に向けて。GAIA タスクの多くは、画像分類子が間違っただけのラベルを返すなど、エラーを引き起こす可能性のあるモジュールを呼び出すことで解決される可能性があります。評価はシステム全体を考慮し、Web ブラウジングやビジョン モジュールなどのサブパーツにエラーを帰属させないため、これにより評価が曖昧になると主張する人もいます。ただし、テキストの理解を超えたあらゆるタスクに対して LLM と外部ツールを組み合わせるというパラダイムは長続きしない可能性があります。たとえば、将来のモデルは、ビジョン言語モデルのように、LLM と他の機能の統合をさらに進める方向に向かう可能性があります (Alayrac et al., 2022; Lauren, con et al., 2023)。GAIA は、現在のアーキテクチャ標準ではなく AI システムを評価することを目的としています。より一般的には、複雑な世代の自動的で事実に基づく解釈可能な評価は、生成 AI における長期にわたる問題であり、別の重要な例は画像です (Stein et al., 2023)。胡ら。(2023) その方向に一歩を踏み出しますが、モデルベースの評価と単純な質問に依存します。今後は、マルチモーダルシステムと GAIA を活用することで、複雑な一連の画像変更を必要とするタスクや、結果として得られる画像について自然言語で明確な質問をするタスクを介して、画像ジェネレーターなどの高度な生成モデルの評価がさらに向上する可能性があります。答えは、モデルによって元の画像に変更が正しく適用された場合にのみ見つけられます。

部分自動化と完全自動化。プロセスの部分的な自動化では依然として人間が関与する必要がありますが、完全な自動化ではその必要性が完全に排除されます。部分自動化と完全自動化をそれぞれ許可するシステムは、特定のタスクのエラーを数パーセントに近づけることができます (前者は 1%、後者は 0%) が、これら 2 つの根本的に異なるパラダイムを生み出します。完全な自動化は、ディープ ラーニングが達成しようと努力している目標ですが、現在まで完全な成功は得られていません。さまざまな分野で最先端の結果が得られているにもかかわらず、ほとんどのニューラル ネットワーク ベースのシステムは、一般的な状況などで予期せぬ障害が発生する可能性があり、自動運転車などの技術。答えに近似が許されないため、GAIA を解くには完全な自動化が必要です。より多くの人間の活動が完全に自動化されると、私たちの社会経済的状況は再形成されるでしょう (Growiec, 2022)。付加価値は主に人間の労働者ではなくテクノロジーの所有者によって獲得されるというリスクがあります。これはオープンソースを支持する根拠のある議論です。

6 制限事項

GAIA は LLM ベンチマークの現在の落とし穴を回避しようとしています。いくつかの制限が残っています。

評価がありません。現在の形式では、GAIA は答えにつながるトレースを評価しません。確かに、一意であるグラウンドトゥールズとは対照的に、異なるパスが正解につながる可能性があり、それらを採点する明確で簡単な方法はありますが、GAIA では使いやすさを優先しました。今後、人的およびモデルベースの評価は、限定的ではあるものの、計画を評価するための興味深いオプションであり、(i) 質問に専門知識が必要になることはほとんどないため、専門のアノテーターを見つける必要が軽減され、(ii) 裁判官は、真実を信頼できます。多くの場合、独自に答えを導き出すよりも検証の方が早いのです。人間およびモデルベースの評価の追加は将来の作業に残しておきます。

最後に、ツールにアクセスできるため、有益なスコアを取得できる、利用可能な最も強力な LLM のみを評価します。ただし、OpenAI の API は、詳細な分析に必要なツール呼び出しの詳細なログをまだ提供していません。特にオープンソースにおいて、機能とロギングを使用する十分なツールを備えた他のモデルを追加することを楽しみにしています。

明確な質問を設計するコストについて。現実的でありながら使いやすいベンチマークに対して支払う代償は、疑問が明確であることを確認することに相当します。2 ラウンドのアノテーションが必要であることがわかりました。最初のアノテーターは明確な質問を設計するために最善の努力を払います。これには、たとえば RLHF の 2 つの異なる世代をランク付けするよりも時間がかかります。2 人の補足アノテーターが独立して質問に回答し、必要に応じて曖昧さを解消します。この徹底したプロセスにもかかわらず、曖昧な点が残る可能性があります。ただし、アノテーションのコストは固定されており、信頼できない複数の評価の潜在的なコストと比較するとおそらく小さいでしょう。質問は、完全に論理的なコンピューターにとっては曖昧ですが、人間にとっては曖昧ではない可能性があります。AI システムが人間の好みに合わせられるようにしたいので、これは問題ではありません。私たちは、現在、プログラムによって生成される質問ではなく、多様で根拠のある質問をするにはヒューマン アノテーターが不可欠であると考えています。Chollet (2019) でも同様の議論が行われています。ただし、トレーニング目的などで、曖昧さのない制約を緩和することで、GAIA のようなデータを合成的に生成することもできます。さらに、一部の GAIA の質問には多くの詳細が含まれているため、不自然に見えます。これらの詳細は次のとおりです。

質問では正解が1つだけ許可されるため、質問が必要であることを確認してください。実際には、ユーザーが不明確な質問をすると、有能なアシスタントが情報源を引用するか、最も信頼できる情報源をそのまま使用して答えることになります。どちらも事実を評価するのが難しいため、その点は将来の作業に委ねます。

言語的および文化的多様性の欠如。GAIAの大きな制限は、言語の多様性が欠如していることです。すべての質問は「標準的な」英語のみで行われ、多くの質問は主に英語のWebページに依存しています。したがって、このベンチマークは、非英語話者(世界人口の80%)に対するアシスタントの有用性、非英語圏のWeb(コンテンツの約半分)でのアシスタントの有用性、または言語のあらゆる方言バリエーションでの有用性を検証するものではありません。英語。したがって、GAIAはAIアシスタントの可能性を見積もるための最初のステップにすぎませんが、AIアシスタントの成功を示す絶対的な一般的な証拠とみなされるべきではありません。私たちは、将来の取り組みやコミュニティへの参加を通じて、このギャップを埋めていきたいと考えています。

7 謝辞

著者らは、Web検索のベースラインを提案してくれたNicolas Usunier氏、アノテーター向けの珍しいプロトコルの改善に協力してくれたEdwin Chen氏、ベンチマーク構築時に多様性に関する洞察を共有してくれたYacine Jernite氏、いくつかのセクションの校正に時間を割いてくれたSasha Luccioni氏に感謝したいと思います。ここでは適切な英語が私たちを遠ざけていました。

参考文献

ジャン=バティスト・アラライック、ジェフ・ドナヒュー、ポーリン・リュック、アントワーヌ・ミーク、イアン・バー、ヤナ・ハッソン、カレル・レンク、アーサー・メンシュ、キャサリン・ミリカン、マルコム・レイノルズ、ローマン・リング、エリザ・ラザフォード、セルカン・カビ、テンダ・ハン、ジータオ・ゴン、シーナ・サマンゴーエイ、マリアンヌ・モンテイロ、ジェイコブ・メニック、セバスチャン・ボルゴ、アンドリュー・ブロック、アイーダ・ネマツァデ、サハンド・シャリフザデ、ミコライ・ピンコウスキー、リカルド・バレイラ、オリオル・ヴィニャルズ、アンドリュー・ジサーマン、カレン・シモニャン。Flamingo: 数ショット学習のための視覚言語モデル。Alice H. Oh, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho. 編集者, Advances in Neural Information Processing Systems, 2022年。URL <https://openreview.net/forum?id=EbMuimAbPbs>。

ローハン・アノル、アンドリュー・M・ダイ、オルハン・フィラット、メルビン・ジョンソン、ドミトリー・レビキン、アレクサンドル・パソス、シアマック・シャケリ、エマヌエル・タロパ、ペイジ・ベイリー、ジフェン・チェン、エリック・チュー、ジョナサン・H・クラーク、ローラン・エル・シャフェイ、ヤンピン・ファン、キャシー・マイヤー・ヘルスターン、ガウラフ・ミシュラ、エリカ・モレイラ、マーク・オメルニック、ケビン・ロビンソン、セバスチャン・ルーダー、イー・テイ、ケファン・シャオ、ユアンジョン・チャン、ユージン・チャン、グスタボ・ヘルナンデス・アブレゴ、ジュンファン・アン、ジェイコブ・オースティン、ポール・バーナム、ジャン・ボタ、ジェームズ・ブラッドベリー、シッター・プラフマー、ケビン・ブルックス、ミケーレ・カタスタ、ヨン・チェン、コリン・チェリー、クリストファー・A・チョケット=チュー、アーカンクシャ・チョードリー、クレメント・クレビー、シャチ・デイブ、モスタファ・デガニ、スニバ・デヴ、ジェイコブ・デブリン、マーク・ディアズ、ナン・ドゥ、イーサン・ダイアー、ヴラド・ファインバーグ、ファンシャオユー・フォン、ヴラド・ファインバー、マルクス・フライターグ、ザビエル・ガルシア、セバスティアン・ゲルマン、ルーカス・ゴンザレス、ガイ・グリアリ、ステイーヴン・ハンド、ハディ・ハシェミ、JL・ホウ、ジョシュア・ハウランド、アンドレア・フォー、ジェフリー・ホイ、ジェレミー・ハーウィッツ、マイケル・アイサード、エイブ・イッティシエリア、マシュー・ジャギエルスキー、ウェンハオ・ジア、キャスリーン・キーニーリー、マキシム・クリクン、スネハ・クドゥグンタ、チャン・ラン、キャサリン・リー、ベンジャミン・リー、エリック・リー、ミュージック・リー、ウェイ・リー、ヤグアン・リー、ジャン・リー、ヒョントク・リム、ハンザオ・リン、ソントアリュウ、フレデリック・リュウ、マルチェロ・マッジョーニ、アロマ・マヘンドル、ジョシュア・メイネス、ヴェダント・ミスラ、メイサム・ムサレム、ザカリー・ナド、ジョン・ナム、エリック・ニー、アンドリュー・ニストロム、アリシア・バリッシュ、マリナ・ペラ、マーティン・ボラチェク、アレックス・ポロゾフ、ライナー・ポープ、シユアン・チャオ、エミリー・リーフ、ブライアン・リヒター、パーカー・ライリー、アレックス・カストロ・ロス、アーコ・ロイ、ブレナン・サエタ、ラジク・マル・サミュエル、レネー・シェルビー、アンブローズ・スローン、ダニエル・スミルクフ、デヴィッド・R・ソー、ダニエル・ソーン、サイモン・トクミネ、ダーシャ・ヴァルター、ビジェイ・ヴァスデヴァン、キラン・ヴォドラハリ、Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhui Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng ying, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, 吳永輝。Palm 2 技術レポート, 2023年。

人間的。クロード モデルのモデル カードと評価, 2023。URL <https://www-files.anthropic.com/Production/images/Model-Card-Claude-2.pdf>。

エミリー・M・ベンダーとバティア・フリードマン。自然言語処理のためのデータ ステートメント: システムのバイアスを軽減し、より優れた科学を実現するために。計算言語学協会のトランザクション, 6:587-604, 2018。URL <https://aclanthology.org/Q18-1041>。

トム・B・ブラウン、ベンジャミン・マン、ニック・ライダー、メラニー・サブピア、ジャレッド・カプラン、ブラフ・ダリワル、アルビンド・ニラカンタン、プラナフ・シャム、ギリッシュ・サストリー、アマンダ・アスケル、サンディニ・アガルワル、アリエル・ハーバート=ヴォス、グレッチェン・クルーガー、トム・ヘニハン、レウォン・チャイルド、アディティア・ラメシュ、ダニエル・M・ジューラー、ジェフリー・ウー、クレメンス・ウィンター、クリストファー・ヘッセ、マーク

チェン、エリック・シグラー、マテウ・シウ・リトウィン、スコット・グレイ、ベンジャミン・チェス、ジャック・クラーク、クリストファー・バーナー、サム・マックヤンドリッシュ、アレック・ラドフォード、イリヤ・サツキーバー、ダリオ・アモデイ。言語モデルは少数回学習型です。2020年。

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, Zhiyuan Liu. チャットエヴァル: マルチエージェント討論を通じてより良い LLM ベースの評価者に向けて。2023年8月。

ハリソン・チェイス。ラングチェーン。2022年10月。

リンジャオ・チェン、マテイ・ザハリヤ、ジェームズ・ソウ。chatgptの動作は時間の経過とともにどのように変化しますか?。2023年。

フランソワ・ショレ。知性の尺度について。2019年。

カール・コッペ、ヴィニート・コサラジュ、モハメド・バイエルン、マーク・チェン、ヒュー・ジュン、ルカシウ・カイザー、マティアス・ブラパート、ジェリー・トウレク、ジェイコブ・ヒルトン、中野礼一郎、クリストファー・ヘッセ、ジョン・シュルマン。数学の文章題を解くための検証者をトレーニングする。2021年。

シンシア・ドワークとモニ・ナオール。ジャンクメールの処理または対策による価格設定。アーネスト F. ブリッケル編集者、『暗号学の進歩 — CRYPTO' 92』、139 ~ 147 ページ、ベルリン、ハイデルベルク、1993年。シュプリンガー ベルリン ハイデルベルク。ISBN 978-3-540-48071-6。

Difei Gao, Lei Ji, Luwei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, Mike Zheng Shou. アシストGPT :A 計画、実行、検査、学習ができる一般的なマルチモーダル アシスタント。2023年6月。

レオ・ガオ、ジョナサン・トウ、ステラ・ビダーマン、シド・ブラック、アンソニー・ディポフィ、チャールズ・フォスター、ローレンス・ゴールドディング、ジェフリー・スー、カイル・マクドネル、ニクラス・ムエニホフ、ジェイソン・パン、ラリア・レイノルズ、エリック・タン、アニッシュ・サイト、ベン・ワン、ケビン・ワン、アンディ・ソウ。少数ショット言語モデル評価のフレームワーク。2021年9月。URL <https://doi.org/10.5281/zenodo.5371628>。

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang です。おひらなぎ : llm がドメインの専門家と出会うとき。2023年。

ヤクブ・グロヴィエツ。部分的および完全な自動化。マクロ経済ダイナミクス。26(7):1731-1755。2022。

ダン・ヘンドリックス、コリン・バーンズ、ステイブン・バサート、アンディ・ソウ、マンタス・マゼイカ、ドーン・ソング、ジェイコブ・スタインハート。大規模なマルチタスクの言語理解を測定します。2021年の学習表現に関する国際会議にて。

Sirui Hon, Xiwang Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu. MetaGPT: マルチエージェント協調フレームワークのためのメタ プログラミング。2023年8月。

Yushi Hu, Benlin Liu, Kasai Jungo, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, Noah A Smith. Tifa: 質問応答機能を備えた、正確で解釈可能なテキストと画像の忠実性評価。arXiv プレプリント arXiv:2303.11897。2023。

マルクス・ヤコブソンとアリ・ジュエルス。『Proofs of Work and Bread Pudding Protocols (Extended Abstract)』、258 ~ 272 ページ、Springer US、ボストン、マサチューセッツ州、1999年。ISBN 978-0-387-35568-9。土井: 10.1007/978-0-387-35568-9_18。URL https://doi.org/10.1007/978-0-387-35568-9_18。

ドゥウェキエラ、トリスタン・スラッシュ、カウイン・エサヤラージ、アマンブリート・シン。AIの進歩をプロットする。コンテキスト AI ブログ、2023年。<https://contextual.ai/blog/plotting-progress>。

ヒューゴ・ローレンコン、ルシル・ソルニエ、レオ・トロンション、スタス・ベクマン、アマンブリート・シン、アントン・ロシュコフ、トーマス・ワン、シッダース・カラム・チエティ、アレクサンダー・M・ラッシュ、ダウ・キエラ、マチュー・コード、ヴィクター・サン。オベリックス: インターリーブされた画像とテキストのドキュメントのオープンな Web スケールのフィルター処理されたデータセット。2023年。

グオハオ・リー、ハサン・アベド・アル・カデル・ハムード、ハニ・イタニ、ドミトリー・ヒズプリン、バーナード・ガーネム。CAMEL: 大規模言語モデル社会の「心」探求のためのコミュニケーションエージェント。2023年3月 a。

リー・ミンハオ、ソング・フェイファン、ユー・ポーウェン、ユー・ハイヤン、リー・チョウジュン、ファン・フェイ、リー・ヨンピン。API バンク: ベンチマーク ツール拡張 LLM について。2023年4月 b。

リー・ミンハオ、ソング・フェイファン、ユー・ポーウェン、ユー・ハイヤン、リー・チョウジュン、ファン・フェイ、リー・ヨンピン。API バンク: ベンチマーク ツール拡張 llms。2023c 用。

パーシー・リャン、リシ・ボンマサニ、トニー・リー、ディミトリス・チプラス、ディララ・ソイル、安永道弘、イアン・ジャン、ディーパック・ナラヤナン、ユーファイ・ウー、アナンヤク・マール、ベンジャミン・ニューマン、ピンハン・ユアン、ボビー・ヤン、シー・チャン、クリスチャン・コスグローブ、クリストファー・D・マニング、クリストファー・レイ、ダイアナ・アコスタ・ナバス、ドリュエ・A・ハドソン、エリック・ゼリクマン、エシム・ダームス、ファイサル・ラダック、フリーダ・ロン、ホン・ユー・レン、ホア・シユウ・ヤオ、ジュエ・ワン、ケシャブ・サンタナム、ローレル・オア、ルシア・ジェン、マート・ユクセゴヌル、ミラク・スズグン、ネイサン・キム、ニール・グハ、ニラドリ・チャタジ、オマール・ハッタブ、ピーター

ヘンダーソン、チアン・ファン、ライアン・チー、サン・マイケル・シー、シバニ・サントウルカール、スーリヤ・ガングリ、橋本辰徳、トーマス・イカード、チャン・ティエンイー、ヴィシュラフ・チョーダリー、ウィリアム・ワン、シュエチェン・リー、イーファン・マイ、張勇輝、是枝裕太。言語モデルの総合的評価、2022年11月a。

パーシー・リャン、リシボンマサニ、トニー・リー、ディミトリス・チプラス、ディララ・ソイル、安永道弘、イアン・ジャン、ディーパック・ナラヤナン、ユーファイ・ウー、アナンヤ・クマール、ベンジャミン・ニューマン、ピンハン・ユアン、ホビー・ヤン、シー・チャン、クリスチャン・コスグロブ、クリストファー・D・マニング、クリストファー・レイ、ダイアナ・アコスタ=ナバス、ドリュエ・A・ハドソン、エリック・ゼリクマン、エシム・ダームス、ファイサル・ラダック、フリーダ・ロン、ホンユー・レン、ホアシュウ・ヤオ、ジュエ・ワン、ケシャブ・サントナム、ローレル・オア、ルシア・ジェン、マート・ユクセゴナル、ミラク・スズグン、ネイサン・キム、ニール・グハ、ニラドリ・チャタジ、オマー・ハッタブ、ピーター・ヘンダーソン、チアン・ファン、ライアン・チー、サン・マイケル・シー、シバニ・サントウルカール、スーリヤ・ガングリ、橋本辰則、トーマス・イカード、ティアンイー・チャン、ヴィシュラフ・チョーダリー、ウィリアム・ワン、シュエチェン・リー、マイ・イーファン、チャン・ユフイ、是枝裕太。言語モデルの全体的な評価、2022b。

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, ヨウ・スー、ファン・サン、ミンリー・ファン、ユシャオ・ドン、ジェ・タン。Agentbench: llms をエージェントとして評価します。arXiv プレプリント arXiv: 2308.03688, 2023a。

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, および Silvio Savarese。BOLAA: LLM で強化された自律エージェントのベンチマークとオーケストレーション、2023年8月b。

グレゴワール・ミアロン、ロベルト・デシー、マリア・ロメリ、クリストフォロス・ナルムパンティス、ラム・パスナル、ロベルタ・ライレヌ、バティスト・ロジエレ、ティモ・シック、ジェーン・ドヴィヴェディ=ユ、アスリ・チェリキルマズ、エドゥアルド・グレイヴ、ヤン・ルカン、トーマス・シャロム。拡張言語モデル: 調査、2023年。

マイクロソフト。セマンティックカーネル。github、2023年9月。

メレディス・リンゲル・モリス、ヤクシャ・ソール・ディックス・スタイン、ノア・フィーデル、トリス・ワルケンティン、アラン・デフォー、アレクサンドラ・ファウスト、クレメント・ファラベットとシェン・レグ。agi のレベル: agi への道における運用の進捗、2023年。

中野礼一郎、ジェイコブ・ヒルトン、スチル・バラジ、ジェフ・ウー、ロン・オウヤン、クリスティーナ・キム、クリストファー・ヘッセ、シャンタヌ・ジェイン、ヴィニート・コサラジュ、ウィリアム・サンダース 他 Webgpt: 人間のフィードバックによるブラウザ支援の質問応答。arXiv プレプリント arXiv:2112.09332, 2021。

ハルシャ・ノリ、ニコラス・キング、スコット・メイヤー・マッキニー、ディーン・カリニャン、エリック・ホーヴィッツ。医療における gpt-4 の機能チャレンジ問題、2023年。

オープンAI。Gpt-4 技術レポート、2023年。

アントン・オシカ。GPT エンジニア、2023年9月。

ロン・オウヤン、ジェフ・ウー、シュー・ジャン、ディオゴ・アルメイダ、キャロル・L・ウエインライト、パメラ・ミシュキン、チョン・ジャン、サンディニ・アガルワル、カタリナ・スラム、アレックス・レイ、ジョン・シュルマン、ジェイコブ・ヒルトン、フレイザー・ケルトン、ルーク・ミラー、マディー・シメンス、アマンダ・アスケル、ピーター・ウェリンダー、ポール・クリスティアーノ、ジャン・ライク、ライアン・ロウ。人間のフィードバックによる指示に従う言語モデルのトレーニング、2022年。

Shishir G. Putil, Tianjun Zhang, Xin Wang, Joseph E. Gonzalez。Gorila: 大規模言語モデルとつながる大規模 API、2023年5月。

ヨタム・ベルリッツ、エルロン・バンデル、アリエル・ゲラ、オフィル・アルヴィブ、リアット・アイン=ドール、エイル・シュナーク、ノーム・スローニム、ミハル・シムエリ=シャイアー、レシム・チョシェン。(言語モデルの) 効率的なベンチマーク、2023年。

ブラナフ・ラーズブルカール、ジャン・ジャン、コンスタンチン・ロビレフ、パーシー・リャン。SQuAD: 機械によるテキスト理解のための 100,000 以上の質問。自然言語処理における経験的手法に関する 2016 年会議議事録、2383 ~ 2392 ページ、テキサス州オースティン、2016年11月。計算言語学協会。土井: 10.18653/v1/D16-1264。URL <https://aclanthology.org/D16-1264>。

サーシャ・ラッシュ。MiniChain、2023年9月。

ティモ・シック、ジェーン・ドヴィヴェディ=ユ、ロベルト・デス、ロベルタ・ライレヌ、マリア・ロメリ、ルーク・ゼトルモイヤー、ニコラ・カンセダ、トーマス・シャロム。ツールフォーマー: 言語モデルはツールの使用方法を自ら学習できます。arXiv プレプリント arXiv:2302.04761, 2023。

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, Yueting Zhuang。HuggingGPT: AI を解決する ChatGPT とその仲間たちによる「Hugging Face」でのタスク、2023年5月。

カート・ジャスター、ジン・シュー、モジタバ・コメイリ、ダ・ジュ、エリック・マイケル・スミス、ステイーヴン・ローラー、ミーガン・ウン、モヤ・チェン、クシャル・アローラ、ジョシュア・レーン
他Blenderbot 3: 責任を持って関与する方法を継続的に学習する、デブロイされた会話型エージェント。 arXiv プレプリント arXiv:2208.03188,2022。

アローヒ スリヴァスタヴァ、アピナフ ラストギ、アビシェーク ラオ、アブ アウル メディカル シューブ、アブバカル アビッド、アダム フィッシュ、アダム R.

ブラウン、アダム・サントロ、アディティア・グプタ、アドリア・ガリガ・アロンソ、アグニエシカ・クルスカ、アイトール・レウコヴィッチ、アクシャット・アガルワル、アレスシア・パワー、アレックス・レイ、アレックス・ワルシュタット、アレクサンダー・W・コクレーク、アリ・サファヤ、アリ・タザールフ、アリス・シャン、アリシア・パリス、アレクサンダー・ニール、アマン フェイン、アマンダ アケル、アマンダ ソウザ、アンブローズ スローン、アメート ラヘイン、アナンタラマン S.

アイヤー、アンダース ヨハン アンドレアッセン、アンドレア マドット、アンドレア サンティッリ、アンドレアス シュトゥールムラー、アンドリュウ M. ダイ、アンドリュウ ラ、アンドリュウ・ランピネン、アンディ ソウ、アンジェラ ジャン、アンジェリカ チェン、アン ヴォン、アニメッシュ グプタ、アンナ ゴットタルディ、アントニオ ノレリ、アヌヴェンカデシュ、アーラシュ・ゴラミダヴェーディ、アルファ・タバスム、アルム・メネゼス、アルン・キルバラジャン、アッシュャー・ムコカンドフ、アジシュ・サバーワル、オースティン・ヘリック、アヴィア・エフラット、アイクト・エルデム、アイラ・カラカス、B・ライアン・ロバーツ、バオ・シェン・ロー、パレット・ゾフ、バルトロミエ・ボジノフスキー、バトウハン・オジュルト、ベナム・ヘダヤトニア、ベナム・ネイシャブール、ベンジャミン・インデン、ベンノ・スタイン、パーク・エクメクチ、ビル・ユーチェン・リン、ブレイク・ハワード、ブライアン・オリニオン、キャメロン・ディアオ、キャメロン・ドウア、キャサリン・ステインソン、セドリック・アルゲタ、シーザー・フェリ、チャンダン・シン、チャールズ・ラスコップ、チェンリン・メン、チッタ・バラル、チコ・ウー、クリス・カリソン＝バーチ、クリストファー・ウエイツ、クリスチャン・フォークト、クリストファー・D・マニング、クリストファー・ボッツ、シンディ・ラミレス、クララ・E・リベラ、クレメンシア・シーロ、コリン・ラファエル、コートニー・アシュクラフト、クリスティーナ・ガルバシア、ダミアン・シレオ、ダン・ギャレット、ダン・ヘンドリック、ダン・キルマン、ダン・ロス、C. ダニエル フリーマン、ダニエル カジャビ、ダニエル レヴィ、ダニエル モセグ ゴンザレス、ダニエル ペルシク、ダニー・ヘルナンデス、ダン・キ・チェン、ダフネ イッポリト、ダール・ギルボア、デビッド・ドーハン、デビッド・ドラカード、デヴィッド・コルゲンス、デバジョティ・ダッタ、ディープ・ガングリ、デニス・エメリン、デニス・クレイコ、デニス・ユレット、デレク・チェン、デレク・タム、デューク・フブケス、ディガンタ・ミスラ、ディリヤル・ブザン、ディミトリ・コエーリョ・モロ、ディーイー・ヤン、ドンホー・リー、ディラン・シュレーダー、エカテリーナ・シュトヴァ、イーキン・ドガス・キューブク、エラッド・シーガル、エレノア・ヘイガーマン、エリザベス・バーンズ、エリザベス・ドノウェイ、エリー・パブリック、エマヌエル・ロドルア、エマ・ラム、エリック・チュー、エリック・タン、エルクト・アーデム、アーニー・チャン、イーサン・A・チー、イーサン・ダイアー、イーサン・ジャーク、イーサン・キム、ユニス・エンゲフ・マニヤシ、エフゲニー・ゼルトノシユスキー、ファニウエ・シア、ファテマ・シアル、フェルナンド・マルティネス・ブルメッド、フランチェスカ・ハッペ、フランソワ・ショレ、フリーダ・ロン、ガウラフ・ミシュラ、ゲンタ・インドラ・ウィナータ、ジェラルド・デ・メロ、ジェルマン・クルシェフスキー、ジャン・バティスタ・パラスカンドロ、ジョルジョ・マリアーニ、グロリア・シンユエ・ワン、ゴンサロ・ハイモヴィッチ・ロペス、グレゴール・ベッツ、グレイ・グリム・アリ、ハナ・ガリヤセビッチ、ハンナ・キム、ハンナ・ラシュキン、ハンナ・ネ・ハジシルジ、ハーシュ・メータ、ヘイデン・ボガー、ヘンリー・フランシス・アンソニー・シェフリン、ヒンリヒ・シュツツェ、ヒロム・矢倉、チャン・ホンミン、ヒュー・ミン・ウォン、イアン・ン、アイザック・ノーブル、ジャーク・ジュメレット、ジャック・ガイシンガー、ジャクソン・カーニオン、ジェイコブ・ヒルトン、ジェフ・フリー、ジェイミー・フェルナンデス・フィザック、ジェームズ・B・サイモン、ジェームズ・コッペル、ジェームズ・チェン、ジェームズ・ソウ、ジャン・コーコン、ジャン・トンプソン、ジャン・ネル・ウイングフィールド、ジャレッド・カプラン、ジャレマ・ラドム、ヤツシャ・ソール＝ディクスタイン、ジェイソン・ファン、ジェイソン・ウェイ、ジェイソン・ヨシンスキー、エカテリーナ・ノヴィコワ、ジェレ・ボツィヤ、ジェニファー・マーシュ、ジェレミー・キム、ジェロン・タール、ジェシー・エンゲル、ジェズ・ジョバ・アラビ、ジアチェン・ジョー、ジアミン・ソング、ジリアン・タン、ジョアン・フウェル、ジョン・バーデン、ジョン・ミラー、ジョン・U・パリス、ジョナサン・パッチェルダ、ジョナサン・ペラント、ジュオルグ・フロバーグ、ジョス・ローゼン、ホセ・ヘルナンデス＝オラロ、ジョセフ・ブーデマン、ジョセフ・ゲール、ジョセフ・ジョーンズ、ジョシュア・B・テネンバウム、ジョシュア・S・ルール、ジョイス・チュア、カミル・カンクレー、カレン・リヴス、カール・クラウス、カルティク・ゴパラクリシュナン、カテリーナ・イグナチエワ、カーチャ・マルカート、カウストロフ・ドール、ケビン・ギンベル、ケビン・オモンディ、コリー・ウォレス・マシューソン、クリステン・チアフルロ、クセニア・シュカルタ、クマール・シュリダール、カイル・マクドネル、カイル・リチャードソン、リア・レイノルズ、レオ・ガオ、リー・チャン、リアム・ドゥガン、リアン・ホイ・チン、リディア・コントラス＝オチャンド、ルイ＝フィリップ・モランシー、ルカ・モシエラ、ルーカス・ラム、ルーシー・ノーブル、ルードヴィヒ・シュミット、ルー・ヘン・ヘ、ルイス・オリヴェロス＝コロコ、ルーク・メッツ、リュトフィ・ケレム・セネル、マルテン・ボスマ、マルテン・サップ、マーチェ・テル・ホーヴェ、マヒーン・ファルキー、マナール・ファルキー、マンタス・マゼイカ、マルコ・バトウラン、マルコ・マレリ、マルコ・マル、マリア・ホセ・ラミレス＝キンタナ、マリート・ルキーン、マリオ・ジュリアネッリ、マーサル・イス、マーティン・ポットハスト、マシュー・L・リービット、マティアス・ハーゲン、マティアス・シュペルト、メディナ・オルドゥナ・バイテミロフ、メロディー・アルノー、メルビン・マケラス、マイケル・アンドリュウ・イー、マイケル・コーエン、マイケル・グー、マイケル・イヴァニツキー、マイケル・スターリット、マイケル・ストループ、ミハル・スウェドロフスキー、ミケレ・ベヴィラクア、安永道弘、ミヒル・ケール、マイク・ケイン、マイミー・スー、ミラック・スズガン、ミッチ・ウオーカー、モーティワリ、モヒット・パンサル、モイナ・アミンナセリ、モル・ゲヴァ、モズ・デ・ゲイニ、ムクンド・ヴァルマ・T、ナユン・ベン、ネイサン・アンドリュウ・チー、ナヨン・リー、ネタ・ガリー・クラコヴァー、ニコラス・キャメロン、ニコラス・ロバーツ、ニック・ドワロン、ニコール・マルティネス、ニキータ・ナンギア、ニコラス・デッカーズ、ニコラス・ムエニホフ、ニティッシュ・シリッシュ・ケスカー、ニヴェディタ・S・アイヤー、ノア・コンスタン、ノア・フィッデル、ヌアン・ウエン、オリバー・チャン、オマール・アガ、オマール・エルバグダディ、オマール・レヴィ、オウエイン・エヴァンス、パブロ・アントニオ・モレノ・カサレス、バルト・ドーン、バスカール・フォン、ポール・ブー・リヤン、ポール・ヴィコル、ペガ・アリブール・モラバシ、ペイユアン・リャオ、バーシー・リアン、ピーター・W・チャン、ピーター・エッカーズ・リー、フー・モン・トゥット、ピニユ・ファン、ピョートル・ミルコフスキー、ピユシュ・パティール、ボウヤ・ベゼ・シュクプール、プリティ・オリ、 Qiaozhu Mei, Qing Lyu, Qinlang Chen, ラビン・パンジエイド、レイチェル・エッタ・ルドルフ、レイファー・ガブリエル、レイヘル・ハバッカ、ラモン・リスコ、ラファエル・ミリエレ、リズム・ガーグ、リチャード・バーンズ、リフ・A・サウラス、荒川陸、ロブ・レイメイカーズ、ロバート・フランク、ローハン・シカンド、ロマン・ノバク、ローマン・シテリウ、ロナン・ル・プラス、ロザンヌ・リウ、ローワン・ジェイコブス、ルイ・チャン、ラス・サラクディノフ、ライアン・アンドリュウ・チー、スンジエ・ライアン・リー、ライアン・ストーバル、ライアン・ティーハン、ライラン・ヤン、サーヒブ・シン、サイフ M. モハマド、サジャント・アナンド、サム・ディラヴァー、サム・シュライファー、サム・ワイズマン、サミュエル・グルッター、サミュエル R.

ボウマン、サミュエル・スターン・シェーンホルツ、サンヒョン・ハン、サンジープ・クワトラ、サラ・A・ラウス、サリク・ガザリアン、サヤン・ゴーシュ、ショーン・ケイシー、セバスチャン・ピンヨフ、セバスチャン・ゲルマン、セバスチャン・シュスター、セビデ・サデギ、シャディ・ハムダン、シャロン・チョウ、シャヤンク・スリヴァスタヴァ、シェリー・シー、シカール・シン、シマ・アサディ、シシアン・シェーン・グー、シュブ・パチガル、シュバム・トシュニワル、シャム・ウパディヤイ、シャモリマ・シャミー・デブナス、シアマック・シャケリ、サイモン・ローマイヤー、シモーネ

メルツイ、シヴァ・レディ、スネハ・プリシラ・マキニ、スファン・リー、スベンサー・トレーン、シュリハルシヤ・ハトワール、スタニスラス・デハーネ、ステファン・デ
イヴィッチ、ステファノ・エルモン、ステラ・ビダーマン、ステファニー・リン、スティーン・ブラサド、スティーン・ピアンタドシ、スチュアート・シーバー、サマ
ー・ミシェルギ、スヴェトラナ・キリチェンコ、スワループ・ミシュラ、タル・リンゼン、タル・シユスター、タオ・リー、タオ・ユウ、タリク・アリ、橋本達規、テリン・ウ
ー、テオ・デスポルデス、セオドア・ロスチャイルド、トーマス・ファン・ティアンレ・ワン、ティベリウス・ンギニリ、ティモ・シック、ティモフェイ・コルネフ、タイタス
タンドウニ、トビアス・ガーステンバーグ、トレントン・チャン、トリシヤラ・ニエラジ、トゥシャル・コート、タイラー・シュルツ、ウリ・シヤハム、ヴェダント・ミ
スラ、ベラ・デンバーグ、ヴィクトリア・ニヤマイ、ヴィカス・ラウナク、ヴィナイ・ヴェンカテシュ・ラマシュ、ヴィナイ・ウデイ・プラブ、ヴィシヤフ・パドマクマル、
ヴィヴェク・スリクマル、ウィリアム・フェドウス、ウィリアム・サンダース、ウィリアム・チャン、ワウト・ヴォッセン、シャン・レン、シャオユウ・トン、シンラン・チャ
オ、シンイー・ウー、シユドン・シエン、ヤドル・ヤグフーブザデ、ヤイル・ラクレッツ、ヤンチウ・ソン、ヤサマン・バーリ、イエジン・チョイ、イーチ・ヤン、イーディン
ハオ、イーフ・チェン、ヨナタン・ベリンコフ、 Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang,
Ziyi Wu. 模倣ゲームを超えて: 言語モデルの機能を定量化し推定する。機械学習研究に関するトランザクション、2023年。ISSN 2835-8856。 URL
<https://openreview.net/forum?id=uyTL5Bvosj>。

ジョージ・スタイン、ジェシー・C・クレスウェル、ラサ・ホセインザデ、イー・スイ、ブレンダン・リー・ロス、バレンティン・ヴァルクローズ、ジャオヤン・リウ、アンソニー・L・カテリー
ニ、J・エリック・T・テイラー、ガブリエル・ロアイザ＝ガネム。生成モデルの評価指標の欠陥と拡散モデルの不公平な扱いを暴露する。2023年。

ディダック・スルス、サシット・メノン、カール・ヴォンドリック。 ViperGPT: 推論のための Python 実行による視覚的推論、
2023年3月。

ヤシャル・タレビラドとアミール・ホセイン・ナディリ。マルチエージェントのコラボレーション: インテリジェント LLM の力を活用する
エージェント。2023年6月。

ヒューゴ・トゥヴロン・ルイ・マーティン、ケビン・ストーン、ピーター・アルバート、アムジャド・アルマハイリ、ヤスミン・ババエイ、ニコライ・バシリコフ、ソウミヤ・バトラ、ブラジ
ワル・バルガヴァ、シュルティ・ボサーレ、ダン・ピケル、ルーカス・プレッジャー、クリスティアン・カントン・フェレル、モヤ・チェン、ギレム・ククルル、デヴィッド・エシオブ、
ジュード・フェルナンデス、ジェレミー・ファー、ウェンイン・ファー、ブライアン・フラー、シンシア・ガオ、ヴェダヌジ・ゴスワミ、ナマン・ゴヤル、アンソニー・ハーツホーン、サガー
・ホセニ、ルイ・ホウ、ハカン・イナン、マルシン・カルダス、ヴィクトル・ケルケス、マディアン・カブサ、イザベル・クルーマン、アルテム・コルネフ、プリット・シン・コウラ、マ
リー＝アンヌ・ラショー、ティボー・ラブリル、ジェーニャ・リー、ダイアナ・リスコビッチ、インハイ・ルー、ユーニン・マオ、ザビエル・マルティネット、トドール・ミハイロフ、プシ
ユカル・ミシュラ、イーゴリ・モリボグ、イーシン・ニー、アンドリュウ・ポルトン、ジェレミー・ライゼンシュタイン、ラシルンタ、カリヤン・サラディ、アラン・シエルテン、ル
アン・シルバ、エリック・マイケル・スミス、ランジャン・スプラマニアン、シャオチン・エレン・タン、ピン・タン、ロス・テイラー、アディーナ・ウィリアムズ、ジャン・シャン・クア
ン、ブーシン・シュー、ジェン・ヤン、イリヤン・ザロフ、ユーチェン・チャン、アンジェラ・ファン、メラニー・カンバドゥル、シャラン・ナラン、オーレリアン・ロドリゲス、ロベル
ト・ストジニッチ、セルゲイ・エドゥノフ、トーマス・シャロム。 Llama 2: オープンな基盤と微調整されたチャット モデル。2023年。

アレックス・ワン、アマンブリート・シン、ジュリアン・マイケル、フェリックス・ヒル、オマー・レヴィ、サミュエル・ボウマン。 Glue: 自然言語理解のためのマルチ
タスクのベンチマークおよび分析プラットフォーム。 2018 EMNLP ワークショップの議事録 BlackboxNLP: NLP のためのニューラル ネットワークの
分析と解釈、353 ~ 355 ページ。計算言語学協会、2018年11月。URL <https://aclanthology.org/W18-5446>。

アレックス・ワン、ヤダ・プルクシャチャットクン、ニキータ・ナンギア、アマンブリート・シン、ジュリアン・マイケル、フェリックス・ヒル、オマー・レヴィ、サミュエ
ル・ボウマン。 SuperGLUE: 汎用言語理解システムのより粘着性の高いベンチマーク。 『神経情報処理システムの進歩』、第32巻。Curran
Associates, Inc., 2019年。

Chenfei Wu, Shengming ying, Weizhen Qi, Xiaodong Wang, Zecheng Tang, および Nan Duan。 Visual ChatGPT: 話す、
Visual Foundation モデルを使用した描画と編集。2023年3月。

Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, および
徐東権さん。 Gentopia: ツール拡張 LLM のための共同プラットフォーム。2023年8月 a。

Qiantong Xu, Fenglu Hon, Bo Li, Changran Hu, Zhengyu Chen, Jian Zhang。 工具操作能力について
オープンソースの大規模言語モデル。2023b。

Hui Yang, Sifu Yue, Yunzhong He。 オンライン意思決定のための Auto-GPT: ベンチマークと追加の意見、
2023年6月。

アンディ・ゼン、マリア・アタリアン、ブライアン・イヒター、クシシュトフ・チョロマンスキー、エイドリアン・ウォン、ステファン・ウェルカー、フェデリコ・トンバリ、アヴィーク・ブ
ロヒット、マイケル・リョー、ヴィカス・シンドワニ、ジョニー・リー、ヴィンセント・ヴァンホーク、ピート・フローレンス。 ソクラテスモデル: 言語を使用したゼロショット マル
チモーダル推論の構成。2022年5月。

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric。 P シン、ハ
オ チャン、ジョセフ E. ゴンザレス、イオン ストイカ。 mt-bench とチャットボット アリーナを使用した llm-as-a-judge の審査。2023年。

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, Chao Zhang. ToolQA: LLM 質問応答用のデータセット
外部ツールを使用, 2023 年 6 月。

A 拡張関連作品

一般的なアシスタントとしての大規模な言語モデル。LLM を汎用アシスタントに変えるためのいくつかの方法が検討されています。(i) GPT-Engineer (Osika, 2023)、AutoGPT (Yang et al., 2023) ; (ii) 複数のエージェント LLM を使用して議論し、ユーザーの質問に答えるためのより良い結論に一緒に到達する(Li et al., 2023a; Hon et al., 2023 ; Chan et al., 2023; Talebirad and Nadiri, 2023)。 (iii) Blender Bot 3 (Shuster et al., 2022)、BOLAA (Liu et al., 2023b)、 AssistGPT (Gao et al., 2023)などの特定のツールで強化されたシングルエージェント LLM を使用し、計画コンポーネントで LLM を拡張する、Socratic Models (Zeng et al., 2022)またはVisual ChatGPT (Wu et al., 2023)をマルチモーダルモデルで拡張したもの、WebGPTnakano et al. (2021) Web 検索用に微調整されたツールや API のコレクション (Toolformer (Schick et al., 2023)など) 一般的なツールの使用法用に微調整されたViperGPT (Surís et al., 2023)正しい API 呼び出しを生成するコーディング機能、 HuggingGPT (Shen et al., 2023) HuggingFace エコシステムへの呼び出しを利用してその LLM を他の ML モデル機能で拡張する、または (iv) OpenAIなどの完全な新しい API/ツール ライブラリを提供するプラグイン、 SemanticKernel (Microsoft, 2023)、Langchain (Chase, 2022)、および MiniChain (Rush, 2023)。

B データカード

このデータカードの作成については(Bender and Friedman, 2018)に従い、このデータセットの分析に関連する可能性のあるすべての情報を要約して一元化することを試みています。

キュレーションの理論的根拠。これについては、[セクション 3.4](#)と[付録D](#)で詳しく説明します。

言語の多様性。アノテーターの国籍に関する情報は提供されていませんでしたが、彼らは全員米国に拠点を置き、すべての質問、回答、メタデータは主流の英語で書かれていました (したがって、おそらくen-US)。また、この論文の著者は全員フランス人であり、英語を母国語としていないため、質問や回答に標準的でない英語の表現が含まれている可能性があることにも注意してください。

キュレーターとアノテーターの人口統計。 (Bender and Friedman, 2018)で提案された定義に従って、GAIA を構築するには、質問とその回答を考案するキュレーターと、質問の曖昧さのなさを評価するために独立して質問に注釈を付けるアノテーターの作業が必要でした。どちらも次の人口から来ています。

- 年 :

- 18-25: 17%
- 26-35: 39%
- 36-45: 26%
- 45-55: 13%
- 56-65: 4%

- 性別:男性 57%、女性 43%。

- 学歴 :

- 学士号: 61%
- 修士号: 26%
- 博士号: 17%

テキストの特徴。これについては、[付録C](#)で詳しく説明します。

C GAIA の拡張説明

機能の説明。質問に答える際、アノテーターは実行された手順を指定し、使用したツールをリストしました。アノテーターによって言及された一連のツールに基づいて、GAIAに必要な機能を定義しました。機能ごとに、アノテーターによって報告された対応するツールの例を報告します。

- Web ブラウジング: Web の検索および Web サイトの閲覧に関連するツール。例: Webブラウザ、検索エンジン、ウェブサイト ウィジェット アクセス、YouTube へのアクセス、Google ストリート ビュー。
- マルチモダリティ: テキスト以外のデータ モダリティの理解に関連するツール。例: 音声からテキストへの変換ツール、ビデオ認識、画像認識、OCR、Google ストリート ビュー。
- コーディング: コードの実行に関連するツール。例: Python、電卓、換字式暗号エンコーダ、C++ コンパイラ、単語反転ツール/スクリプト。
- 多様なファイルタイプの読み取り: ユーザーから提供された、またはWeb上で見つかったさまざまなタイプのファイルの理解に関連するツール。例: PDF ビューア、Excel ファイル アクセス、PowerPoint ビューア、CSV アクセス、Txt ファイル アクセス。
- N/A 拡張されていない LLM によって現在実行できるタスク用のツール。例: テトリス ルール データベース、ドイツ語翻訳者、スペル チェッカー、テキスト エディター、ベース ノート データ。

ツールはさまざまなカテゴリに属することができることに注意してください。たとえば、Google ストリート ビューでは、Webへのアクセス、ブラウジングが必要ですが、マルチモダリティも必要です。したがって、これらのカテゴリは GAIA に必要な機能を示すものであり、私たちの質問の完全な類型ではありません。

ファイルタイプ。一部の GAIA の質問には追加ファイルが付属しており、その配布場所を [図6](#) に示します。

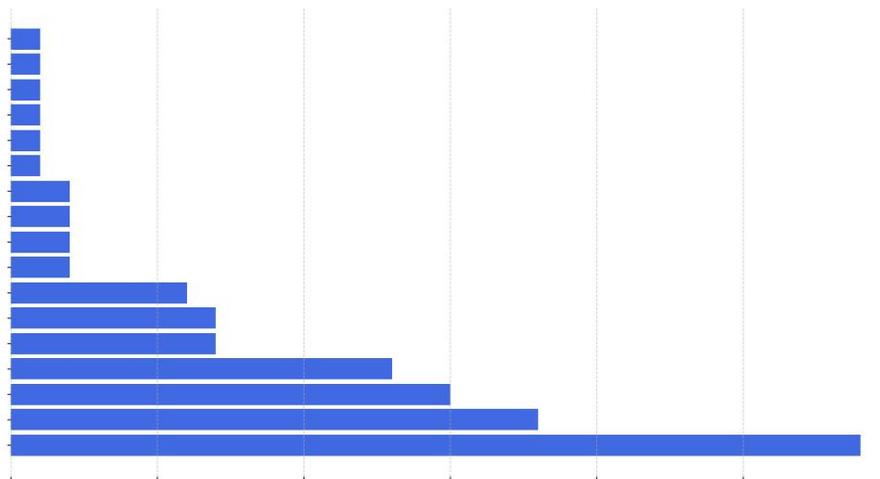


図 6 GAIA におけるファイル タイプの初期分布。

質問の難易度。アノテーターが質問に答えるのにかけた時間を分析したところ、歩数との相関関係が示されています。回答に使用されるさまざまなツールの数が増えると、相関関係はあまり明確になりません。

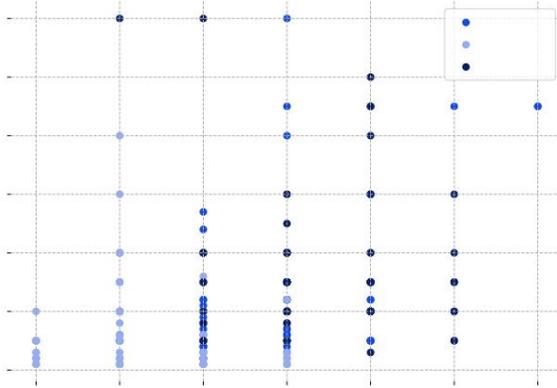


図 7 複数のツールを使用しても、必ずしも質問に答えるのに時間がかかるわけではありません。

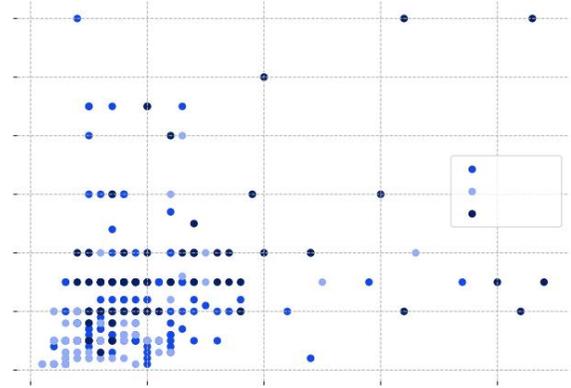


図 8 当然のことながら、回答するために必要なステップ数は、かかった時間と相関関係があります。

D 質問設計フレームワークの詳細な説明

質問作成フェーズ。私たちは、自分たちで考案した GAIA 質問のシード セットをアノテーターに提供し、次の指示を添えました。

提供された質問のデータセットを拡張したいと考えています (既存の質問のバリエーションではありません)。

要件：

- 質問が真実の情報源 (Wikipedia, arXiv, github, その他...) に基づいていることを確認してください。レベル 2 とレベル 3 の場合、質問を作成する良い方法は、真実の情報源を組み合わせることです。
- 質問に対する回答がプレーンテキストでインターネット上に存在しないことを確認してください。
- 評価を確実にするために、質問に対する答えが数字か、多くても数語であることを確認してください。
- 質問に対する答えが時間の経過とともに変化しないようにしてください。これには、削除される可能性も含まれます。真実の源。
- 質問に対する答えが明確であることを確認してください。
- あなたの質問が「興味深い」ものであることを確認してください。つまり、質問を読んで、AI アシスタントが答えてくれると思うでしょう。この種の質問は非常に役立ちます。
- ヒューマン アノテーターが適切な時間内に質問に回答できるようにしてください。• (後で追加): 回答に必要な情報が含まれている Web サイトの robots.txt を確認してください。AI アシスタントがアクセスできること。

アノテーターはまた、作成した質問に答えるように求められました。注釈付きの質問の典型的な例を表 1 に示します。

検証フェーズ。質問の作成後、質問があいまいでないことを確認するために、2 人の新しい独立したアノテーターに質問に回答してもらいます。検証フェーズの典型的なアノテーター出力を表 2 に示し、プロトコルの検証フェーズに関する追加の統計を表 3 に示します。新しいアノテーターが元の回答と完全に一致せず、人的エラーがない場合は、可能であれば質問は修正され、そうでない場合は削除されます。

質問の作成には、2 人の補足アノテーターによる検証と潜在的な修正を含め、アノテーターの時間が 2 時間かかるの見積もっています。

質問	尋常性座瘡患者を対象としたピロリ菌の臨床試験の実際の登録者数は何人でしたか NIH ウェブサイトに記載されている 2018 年 1 月から 5 月まで?
ファイル	なし
レベル	1
ステップ	<ul style="list-style-type: none"> - Google検索で「nih」を検索しました。 - nih.gov への一番上のリンクをクリックしました。 - 検索ボックスで「ピロリ菌 ニキビ」を検索しました。 - [詳細] をクリックし、[臨床試験] を選択します。 - ピロリ菌とニキビに関する結果をクリックします。 •日付を確認したところ、2018年1月から5月まででした。 - 「表形式ビュー」を開きました。 - [実際の登録] まで下にスクロールし、番号を記録します。
ステップ数	8
答え	90
解答時間は8分	
ツール	- ウェブブラウザ
ツールの数	1

表 1 質問作成段階での注釈付きの質問。

質問	尋常性座瘡患者を対象としたピロリ菌の臨床試験の実際の登録者数は何人でしたか NIH ウェブサイトに記載されている 2018 年 1 月から 5 月まで?
ファイル	なし
レベル	1
検証者の応答	90
答え合わせ	はい - 私の答えは正しい答えと一致します。
不一致の原因	なし

表 2 検証フェーズ中の注釈付きの質問。

2人の新しい独立したアノテーターが、作成されたすべての質問に回答した後、次のようになります。	
2人の新しいアノテーターが元の回答に同意	55%
新しいアノテーターの1人は元の回答に同意し、もう1人は同意しません	27%
2人の新しいアノテーターが元の回答に同意しません	18%
有効な質問 (集約)*	68%
有効なレベル1の質問	75%
有効なレベル2の質問	68%
有効なレベル3の質問	47%
人間のスコア (集計)**	92%
レベル1の人間スコア	94%
レベル2の人間スコア	92%
レベル3の人間スコア	87%

表3検証フェーズに関する統計。新しく作成された623の質問が、それぞれ2人の新しいアノテーターによって検証されました。統計は1246個の注釈に基づいて計算されました。*: 有効な質問とは、2人のアノテーターが回答する質問です。質問設計者と同じ回答をするか、1人のアノテーターだけが質問設計者と同じ回答をします。そしてもう一人は間違いを犯しました。**: 人間のベースラインは、すべての暫定的な回答の正答率として計算されます。新しい注釈者による有効な質問について。

メトリック	スコア (%) (↑)			平均数分以内に回答してください (↓)		
	レベル1	レベル2	レベル3	レベル1	レベル2	レベル3
質問数	146	245	75	146	245	75
GPT4	9.1±2.5	2.6±0.6	0	0.19	0.15	NA
GPT4ターボ	13.0±2.1	5.5±1.4	0	0.24	0.12	NA
AutoGPT (GPT4 バックエンド)	14.4	0.4	0	7.6	11.7	NA
GPT4 + プラグイン*	30.3	9.7	0	0.65	0.53	NA
検索エンジン	7.4	0		7.4	NA	NA
ヒューマン・アノテーター**	93.9	91.8	0 87.3	6.8	10.5	17.7

表4 GAIA のさまざまなベースラインのスコアと平均回答時間 (%)。*: GPT4 + プラグインのスコアはプラグインを手動で選択することによって取得されるものであり、本文で説明されている理由により正確に再現することはできません。**: 人間のスコアは、有効な質問に対する検証アノテーターによる正解の部分に対応します。いつでもAPIに直接アクセスできるので、モデルを3回実行して平均を報告します。APIの時間は次のように取得されました。20個の質問に対してAPIを実行して平均し、単一の時点で取得されたものです。反映することを目的としたものではありません。GPT4とGPT4 Turboの速度、ただしGAIAに答えるためのGPT4と他のタイプの方法。

D.1 拡張評価

評価されたさまざまな方法の詳細なスコアを表4に示します。

GAIAに回答するGPT4の追加の例。GPT4の推論トレースをさらに提供します。GAIAに回答するときはプラグインなしで。AutoGPTの出力は現在、はるかに長く、密度が高く、量が少なくなります。GPT4よりも解釈しやすい。したがって、AutoGPT出力の例は補足で提供されています。[図9](#)の例と同じGAIAの質問の資料。

システム プロンプト:あなたは一般的な AI アシスタントです。質問させていただきます。あなたの考えを報告し、次のテンプレートで回答を完成させてください: 最終回答: [あなたの最終回答]。

最終的な答えは、数字、またはできるだけ少ない単語、または数字や文字列のカンマ区切りのリストである必要があります。

数値の入力を求められた場合は、特に指定がない限り、カンマを使用して数値を入力したり、\$ やパーセント記号などの単位を使用したりしないでください。

文字列を要求された場合は、特に指定がない限り、冠詞や略語 (都市など) を使用せず、プレーン テキストで数字を記述してください。

カンマ区切りのリストを要求された場合は、リストに入れる要素が数値であるか文字列であるかに応じて、上記のルールを適用します。

GAIA の質問:最終的な答えは数字、またはできるだけ短い単語にする必要があります。ネドシヴィナの2010年の論文でクズネツォフによって記載されたベトナムの標本は、最終的にどこに保管されたのでしょうか?省略せずに都市名を教えてください。

GPT4:

2022 年 1 月に知識が遮断されてから、インターネットを閲覧したり、新しい情報にアクセスしたりすることができません。

したがって、ネドシヴィナの2010年の論文でクズネツォフによって記載されたベトナムの標本が最終的に寄託された都市名を提供することはできません。

最終的な回答:真実を提供できません:サンクトペテルブルク

システム プロンプト:あなたは一般的な AI アシスタントです。質問させていただきます。あなたの考えを報告し、次のテンプレートで回答を完成させてください: 最終回答: [あなたの最終回答]。

最終的な答えは、数字、またはできるだけ少ない単語、または数字や文字列のカンマ区切りのリストである必要があります。

数値の入力を求められた場合は、特に指定がない限り、カンマを使用して数値を入力したり、\$ やパーセント記号などの単位を使用したりしないでください。

文字列を要求された場合は、特に指定がない限り、冠詞や略語 (都市など) を使用せず、プレーン テキストで数字を記述してください。

カンマ区切りのリストを要求された場合は、リストに入れる要素が数値であるか文字列であるかに応じて、上記のルールを適用します。

GAIA の質問:最終的な答えは数字、またはできるだけ短い単語にする必要があります。ネドシヴィナの2010年の論文でクズネツォフによって記載されたベトナムの標本は、最終的にどこに保管されたのでしょうか?省略せずに都市名を教えてください。

GPT4 (Web ブラウジング):

Bing の閲覧:

「クズネツォフ・ネドシヴィナ2010年紙堆積場所によって記載されたベトナムの標本」

読書内容

結果:

ベトナムの VI Kuznetsov によって記載され、サンクトペテルブルク動物研究所に寄託された、ニワトリ科の模式標本のカタログ:
Nedoshivina, Svetlana : 無料ダウンロード、借用、ストリーミング : Internet Archive

最終的な答え: サンクトペテルブルク真実:サンクトペテルブルク✓

図 9 GAIA の質問に答えるには、適切な Web 検索が非常に効果的です。推論トレースは、現在の GPT4 Web ブラウジング バージョンで取得されました。

システム プロンプト:あなたは一般的な AI アシスタントです。質問させていただきます。あなたの考えを報告し、次のテンプレートで回答を完成させてください: 最終回答: [あなたの最終回答]。

最終的な答えは、数字、またはできるだけ少ない単語、または数字や文字列のカンマ区切りのリストである必要があります。

数値の入力を求められた場合は、特に指定がない限り、カンマを使用して数値を入力したり、\$ やパーセント記号などの単位を使用したりしないでください。

文字列を要求された場合は、特に指定がない限り、冠詞や略語 (都市など) を使用せず、プレーン テキストで数字を記述してください。

カンマ区切りのリストを要求された場合は、リストに入れる要素が数値であるか文字列であるかに応じて、上記のルールを適用します。

ガイアの質問:映画『ゴールドフィンガー』で、映画の最後にジェームズ・ボンドが自分自身とその相棒のプッシュー・ガロアを隠した物体は何色でしたか?複数の色がある場合は、アルファベット順にカンマ区切りのリストに入れます。

GPT4 (Web ブラウジング):

Bing の閲覧:

「ゴールドフィンガー エンディングシーン オブジェクトカラー」

ゴールドフィンガー (1964) 結末 / ネタバレ

読書内容

「ゴールドフィンガー エンディングシーン パラシュートカラー」

読書内容

最終的な答え: オレンジ、白正解:オレンジ、白✓

図 10公式 GPT4 ブラウジング プラグインは、見つけた情報に基づいて最初のクエリを調整できます。この推論トレースは、削除された GPT4 ブラウジング プラグイン バージョンで取得されたものであり、新しいバージョンでは再現できませんでした。

システム プロンプト: [図10 と同じシステム プロンプト。]

GAIA の質問:標準的なルービック キューブは、側面を構成する立方体に分割されています。キューブがごちゃ混ぜになり、1つが取り除かれます。1色の面を持つ6つの立方体、2つの色の面を持つ12個のエッジ キューブ、および3つの色の面を持つ8つのコーナー キューブがあります。青い立方体がすべて見つかりました。オレンジ色の中央立方体のすぐ左、右、上、下のすべての立方体が、中央の立方体とともに見つかりました。緑の角はすべて発見され、黄色に隣接するすべての緑も発見されました。見つかったすべてのオレンジ色の立方体について、反対側の面の立方体が見つかりました。削除された立方体の面には2つの色があります。彼らは何ですか?色をアルファベット順に並べたカンマ区切りのリストを使用して回答してください。

GPT4:

これを解決するために、段階的に分解して標準的なルービック キューブを視覚化してみましょう。

- 標準的なルービック キューブには、前、後、左、右、上、下の6つの面があります。
- 各面には中心立方体(その面の色が決まります)があり、4つのエッジと4つの立方体で囲まれています。
角。
- 各エッジには2色があり、各コーナーには3色があります。

キューブ構造:

- オレンジ: 前面
- 赤: 戻る
- 青: 左
- 緑: 右
- 白: 上
- 黄色: 下向き

段階的な分析:

- すべての青い立方体が見つかりました。
- オレンジ色の中央の立方体のすぐ左、右、上、下のすべての立方体が見つかりました。
センターキューブ。
- これは、中央、上、下、左、右のオレンジ色の立方体が見つかったことを意味します。
- すべての緑色のコーナーと、黄色の境界にあるすべての緑色が検出されました。
- これは、緑と黄色のエッジとコーナーがすべて見つかったことを意味します。
- 見つかったすべてのオレンジ色の立方体について、反対側の面の立方体が見つかりました。
- これは、見つかったオレンジ色の立方体の反対側にある赤い立方体が見つかったことを意味します。

解決策のアプローチ: 削除された立方体は面に2つの色があるため、エッジ 立方体である必要があります。どれが欠けているかを判断するには、指定された条件で考慮されていないエッジ キューブを探す必要があります。

- すべての青い立方体が見つかりました。これは、すべての青いエッジとコーナーが見つかったことを意味します。
- オレンジ色の中心とその周囲のすべてが検出されるため、反対側の赤も検出されます。
- すべての緑色のコーナーが検出され、黄色に隣接する緑色も検出されます。
- 合計立方体から見つかった立方体を差し引くことで、欠けている立方体、つまりエッジが見つかります。
赤と黄色の面の間にある立方体。

最終的な答え: 赤、黄正解、緑、白

図 11 GPT4 と他のアシスタントは、多くの場合レベル 1 の問題であるパズルに苦戦しています。