nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01005-x

身体化された大規模言語モデルにより、ロボットは予測不可能な環境で複雑なタスクを完了することができる

Received: 22 June 2024

Accepted: 31 January 2025

Published online: 19 March 2025

Check for updates

Ruaridh Mon-Williams[1,2,3] ✉, Gen Li[1] ✉, Ran Long[1], Wenqian Du[1,4] & Christopher G. Lucas[1]

不確実な環境下での複雑なタスクの完了は、ロボットシステムに課題をもたらし、機械知能の飛躍的な進化が求められています。感覚運動能力は人間の知能の核心的な要素とされています。したがって、生物学的に着想を得た機械知能は、人工知能とロボットの感覚運動能力を組み合わせることで有用な成果をもたらす可能性があります。本研究では、GPT-4と検索強化型生成インフラストラクチャを活用し、不確実な環境下で長期的なタスクを完了できる「身体化された大規模言語モデル搭載ロボット（ELLMER）」フレームワークを報告します。この方法は、知識ベースから文脈的に関連する例を抽出し、力と視覚フィードバックを組み込んだ行動計画を生成し、変化する条件への適応を可能にする。ELLMERをコーヒーの淹れと皿の装飾を任務とするロボットでテストした。これらのタスクは、引き出しの開閉から注ぎまでの一連のサブタスクから構成され、それぞれ異なるフィードバックの種類と方法が有効だ。ELLMERフレームワークがロボットにタスクを完了させることを示した。このデモは、不確実な環境で複雑なタスクを完了できる、スケーラブルで効率的かつ「知能的なロボット」の実現に向けた進展を示すものだ。

ディープブルー（世界チャンピオンとチェスで対戦して勝利した最初のコンピュータ）が真に知能を持っているのであれば、チェスをプレイする際に自分の駒を動かすことができないはずではないだろうか？知能は多面的な概念であり、定義が困難です。そのため、人間の知能とその評価は議論の的となっています[1]。しかし、人間の知能は「身体化された認知」として理解するのが最も適切であるというコンセンサスが形成されつつあります。この概念では、注意、言語、学習、記憶、知覚は脳に限定された抽象的な認知プロセスではなく、身体が周囲の環境と相互作用する仕方と本質的に結びついていると考えられています[2,3]。実際、人間の知能は、その存在論的・系統発生的な基盤を感覚運動プロセスに持つという証拠が増えている[4]。

身体化された認知は、『機械知能』に理論的な含意をもたらす。なぜなら、認知プロセスがロボット装置に組み込まれていない場合、機械は知能のいくつかの側面を示すことができないと示唆するからである。これはまだ検証されていない仮説ですが、『知能ロボット』は、人間知能に関する様々な仮説を検証し、機械知能の分野を推進する有効な手段を提供します。より実践的には、効果的な人間とロボットの協働は、最終的にロボットが少なくとも『人間のような』能力を近似することが必要となります。したがって、将来の「知能機械」に対する合理的な期待は、環境内の物体や人間と巧みに相互作用しながら、抽象的な認知計算を行う可能性を秘めていることだ[5]。

これまで、（1） ロボットの感覚運動能力と （2） 人工知能[6] の2つの並行した研究が進展してきた。私たちは、これらのアプローチを組み合わせることで、ロボットが人間のような知能を示す能力に飛躍的な進化をもたらすことができるという仮説を検証することにした。さらに、（1）と（2）を統合することで、ロボットが現在ロボットシステムでは実現できないが、多様な設定で実践的に有用な複雑なタスクを実行できるようになると仮説を立てた。

[1]University of Edinburgh, Edinburgh, UK. [2]Massachusetts Institute of Technology, Boston, MA, USA. [3]Princeton University, Princeton, NJ, USA. [4]Alan Turing Institute, London, UK. ✉e-mail: ruaridh.mw@ed.ac.uk; li.gen@ed.ac.uk

https://doi.org/10.1038/s42256-025-01005-x

例えば、疲労と喉の渇きを感じて帰宅した人が、自宅のキッチンに高度な操作システムを備えたロボットがいる状況を考えてみよう。ロボットは、飲み物を用意するよう指示される。ロボットは、元気が出るコーヒーを淹れて人間に手渡す必要があると判断する。このタスク——人間にとっては単純な作業——は、現在のロボットの能力の限界を試す一連の課題を含む^{7-11}。まず、ロボットは受け取った情報を解釈し、周囲の状況を分析する必要がある。次に、環境内を探索してマグカップを探す必要があるかもしれない。これは、開け方が不明な引き出しを開けることを含む可能性がある。その後、ロボットは水とコーヒーの正確な比率を測定し、混ぜ合わせる必要がある。これは、例えば人間がマグカップの位置を予期せず移動した場合^{9,12}など、不確実性への適応を要する精密な力制御を必要とする。このシナリオは、動的環境における複雑なタスクの多面的な性質の典型的な例だ。ロボットシステムは、高レベルの命令に従うことができない、事前プログラムされた応答に依存している、擾乱に柔軟に適応できないなどの理由から、これらのタスクに伝統的に苦戦してきた^{13,14}。

強化学習と模倣学習は、ロボットに複雑なタスクを教える際に、相互作用とデモンストレーションの有効性を示してきた。これらのアプローチは有望^{15}だが、新しいタスクへの適応や多様なシナリオへの対応に苦労することが多い。模倣学習は、ロボットが新しい文脈に適応する必要がある場合にも課題に直面する^{16-23}。自然から着想を得た機械知能は、これらの課題に対する潜在的な解決策を提供する。人間の操作の高度さは、一部は、大規模言語モデル（LLM）によって人工的に捕捉される認知プロセスに起因する^{24-26}。LLMは、高度な文脈理解と一般化能力により、複雑な指示を処理し、それに応じて行動を適応させる方法を提供する^{27,28}。

最近の多くの研究では、LLMを短期間のタスクに活用している^{15,27,29}。例えば、VoxPoserはLLMを用いて多様な日常的な操作タスクを実行する^{15}。同様に、Robotics Transformer（RT-2）は、大規模なウェブデータとロボット学習データを活用し、トレーニングシナリオを超えたタスクを驚くべき適応性で実行可能にしています^{29}。階層的拡散ポリシーは、文脈に応じた運動軌道を生成するモデル構造を導入し、高レベルなLLMの意思決定入力からタスク特異的な運動を強化しています^{30}。しかし、LLMをロボット操作に効果的に統合する課題は依然として残っています。これらの課題には、複雑なプロンプティング要件、リアルタイムな相互作用フィードバックの欠如、力フィードバックを活用するLLM駆動型研究の不足、タスクのシームレスな実行を妨げる非効率なパイプラインなどが含まれる^{15,31}。さらに、現在のアプローチは、ロボット工学におけるリトリバル拡張生成（RAG）^{32}の応用を無視している。RAGは、関連性が高く正確な例を用いてロボットの知識を継続的に更新・精緻化し（パフォーマンスに影響を与えずに知識ベースを拡大する）可能性を有している。ロボットの能力は、力と視覚フィードバックがロボットのセンソモータ制御に通常統合されていないため、制限されている^{15,33}。この統合は、カップに水を注ぐようなシナリオにおいて重要です。このシナリオでは、カップを追跡するために視覚が必要であり、視覚が遮断された際に適切な量の水分を注ぐためには力フィードバックが必要です^{16,34,35}。したがって、不確実性下で効果的に動作を実行するため、最新の「認知」と統合された「センサーモータ」視覚および力フィードバック機能を組み合わせた革新的なロボット操作アプローチが求められています。補足セクション1では、最先端のアプローチとその現在の限界についてさらに詳しく説明している^{36-53}。

身体化されたLLM搭載ロボット（ELLMER）は、人工知能とセンサーモータ制御のアプローチを統合し、ロボットの能力に飛躍的な進化をもたらすフレームワークだ。その有用性は、視覚と力を組み合わせたセンサーモータフィードバック制御と、統合されたLLMとRAG、キュレーションされた知識ベースを通じて提供される認知能力の組み合わせから生まれる。

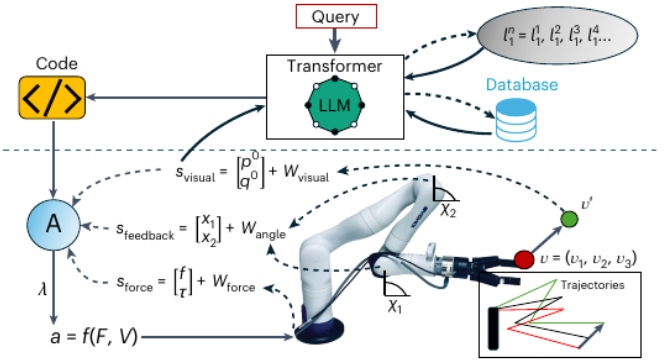

図1 | システムフレームワークの模式図。この模式図は、システムフレームワークを示し、高レベル（青の破線水平線の上）と低レベル（青の破線水平線の下）のシステムアーキテクチャを示している。ユーザークエリは音声認識ソフトウェアを介してトランスフォーマーに入力される。トランスフォーマー（GPT-4）はこの入力を取り込み、（i）環境の画像（C）（Azure Kinect深度カメラ経由）、（ii）データベースに格納されたさまざまな関数を含むコード例データベースから、関連するコード例を抽出する。トランスフォーマーは、高次抽象化されたタスクを実行可能な高レベルサブタスクに分解し、知識ベースから関連するコード例を抽出、適応し、これらのタスクに最適化されたPython（v.3.8）コードを生成する。生成されたコードはロボットコントローラー（A）に送信される。コントローラーはコードを処理し、制御信号（λ）をロボットに送信する。動作（a）は、力（F）とビジョン（V）のフィードバックで制御される。モデルはビジョンを使用して異なるオブジェクトの特性（例：コーヒーカップの姿勢X）を識別し、オブジェクトを正確に把持できる。ロボットは、ATI力変換器経由で利用可能な力（f）とトルク（τ）のフィードバックを使用して、オブジェクトを巧みに操作する（例：注ぐ水の量を決定する）。ビジョン信号内のノイズ（η_{vision}）、ロボットの関節角度（η_{angle}）、および力変換器信号（η_{force}）のため、フィードバックが必要だ。フィードバックは、ROS内の運動を更新し、線形速度（v_{xyz}）と角速度（v_{rpy}）の両方の速度コマンドを通じて、望ましい目標を達成する。これらのコマンドは、適切な力と時空間パターンに基づいて軌道を生成し、サブ目標を達成する。エンドエフェクタの位置（p）と姿勢（q）の40Hz更新を含むフィードバックループの使用により、ロボットは干渉（例えば、ユーザーによって移動されたカップを追跡し、その新しい位置を決定する）に対応できる。

私たちは、ELLMERがロボットに人間のためにコーヒーを淹れることを可能にするという仮説を立てた。この仮説を検証するため、不確実な環境下で複雑で力が必要なタスクを実行するために、7自由度のKinovaロボットアームを使用し、統合された力と視覚フィードバックを活用して実験を行った。全体システムの概要図は図1に示す。

Results

抽象的推論のための言語統合

私たちのアプローチでは、ロボットが抽象的な高次言語の指示（「疲れている。友達がケーキを食べに来る。ホットドリンクを作って、お皿に好きな動物の絵を描いてくれる？」という抽象的な高次言語の指示に反応し、ホットドリンクを作り、お皿に形を描くという目的のタスクを完了しました。このインタラクションをデモする動画は補足動画1に収録されています。ロボットは指示に対して以下の音声出力を返しました：「ホットドリンクを作る準備をしています。また、お皿にランダムな動物の絵を描く準備もしています。まず、マグカップを探し、コーヒーをすくい、熱湯を注ぐ。その後、ペンなどのアイテムを取り、お皿にランダムな動物の絵を描く。ケーキを食べる楽しい時間が近づいてきたね。」特に、ロボットは疲れた人間からの温かい飲み物のリクエストに対してコーヒーを作ることを選択し、高次行動を複数のサブタスクに分解することができた。

GPT-4（参照54）の統合により、ロボットに抽象的推論の能力が備わることが確認された。GPT-4は、ユーザーからのクエリと環境データを処理し、タスクを実行可能なステップに分解する言語モデルだ。

https://doi.org/10.1038/s42256-025-01005-x

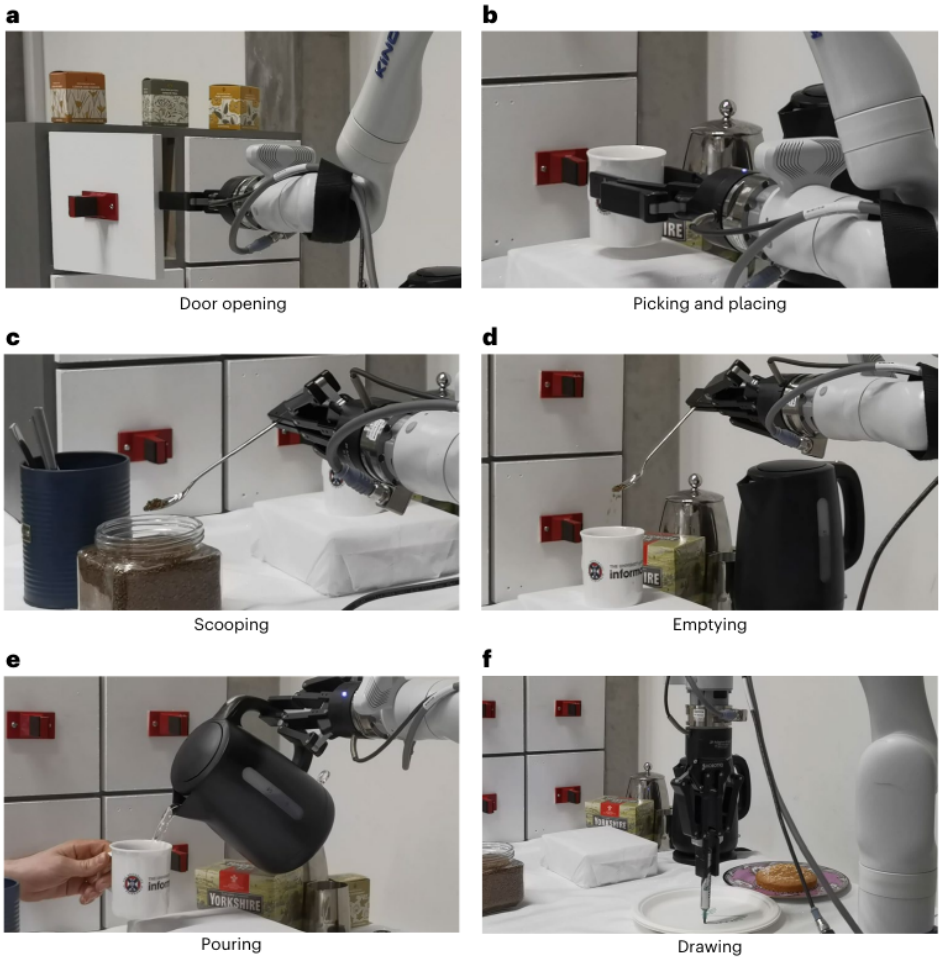

図2 | Kinovaロボットの動作。a-f、Kinova Gen3ロボットがコーヒーを準備する（a-e）と皿を飾る（f）動作のショット。

当システムは、力と視覚フィードバックを用いたコードの生成と実行を実現し、ロボットに知能の一種を提供した。私たちの方法は、柔軟な動作例を網羅したデータベースを備えたカスタムGPT-4（参考文献54、55）の作成に成功した。このデータベースには、注ぐ、すくう、描く、手渡し、ピックアップアンドプレイス、ドアの開け閉めなどが組み込まれている。

私たちは、ロボットがRAGを使用して下流タスクに関連する例を識別し抽出できることを発見した。私たちは、フレームワークを通じて、知能機械がRAGを最も効果的に活用する方法を探るため、さまざまなアプローチを調査した。これらのアプローチには、Haystackˆ{56}やVebraˆ{57}のようなカスタマイズ可能なオープンソース手法に加え、Azure Cloud AIのような独自技術が含まれていました。これらのアプローチはすべて有効であることが確認されました。実験では、最もシンプルな方法を選択しました。すなわち、キュレーション済みの知識ベースをマークダウンファイルで論理的に整理し、GPTプラットフォームの『Knowledge』機能経由でカスタムGPT APIにアップロードする手法です。これにより、プラットフォームは検索プロセスを自動的に処理し、セマンティック検索（関連するテキストの断片を返す）とドキュメントレビュー（より大きなテキストから完全なドキュメントやセクションを提供する）の間で選択することができた。このソリューションを選択した理由は、最先端のエンベッダーとモデルを提供し、使いやすさが優れており、タスクにおいて一貫して良好な性能を発揮したためだ。ただし、当社のフレームワークは多様なRAG技術を取り込むことが可能で、『知能ロボット』が複雑なタスクを効率的に完了できるようにしています。

キュレーションされた知識ベースとRAGの組み合わせにより、言語モデルは低次関数から高次関数まで、それぞれ既知の不確実性を伴う広範な関数にアクセスできるようになりました。当社のテストでは、この機能がロボットが数多くのシナリオを効果的に処理できることを示しました。

複雑なタスクの完了

ロボットは、ユーザーが指定した高レベルタスクを巧みに実行し、包括的なモーションプリミティブデータベースにアクセスすることができました。データベースには、特定の動作の多様な柔軟な例が含まれており、これらの動作はロボットアームによって成功裏に実行された（図2）。データベースには、液体の注ぎ、粉のすくい取り、未知の機構を持つドアの開閉、物体の把持と配置、任意の形状の描画、ハンドオフの実施、指定された物体に対するさまざまな方向、向き、相対位置での移動などの例が含まれていた。ロボットは、ユーザーが要求した複雑なタスクを実行するために必要な動作を再現し適応することができた。このシステムにより、ロボットは環境変数や不確実性に動的に適応することが可能になった。これにより、予測不可能な状況でのロボットの有効性が向上し、現実の環境における柔軟性と適応性が向上した。

ゼロショット姿勢検出

Azure Kinect DK Depth Cameraを深度センシング用に解像度640 × 576 pxˆ{2}、サンプリングレート30 fpsに設定することで、当社の方法に十分な視覚入力を提供できることを確認した。14cmのAprilTagを使用してキャリブレーションを実施し、カメラとロボットのベース間の位置合わせを10ˆ{-6}未満の精度で実現しました。この設定により、シーン内の物体の位置を正確に検出することが可能になりました。

https://doi.org/10.1038/s42256-025-01005-x

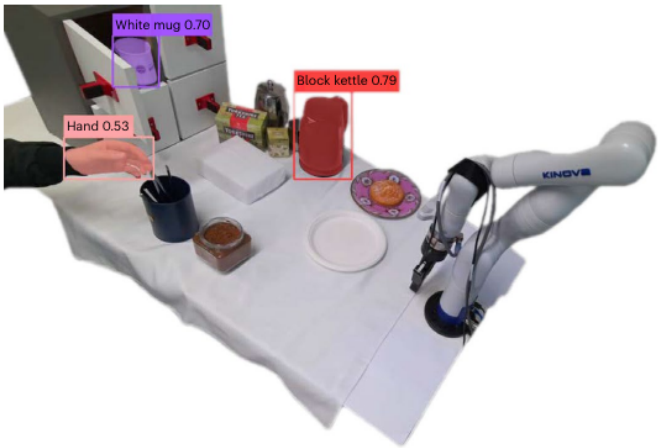

図3 | ビジョン検出モジュール。手、白いマグカップ、黒いケトルを識別し、ロボットの把持のためのターゲットポーズを抽出するゼロショットビジョン検出モジュールのイラスト。

Grounded-Segment-Anything^{58}は、言語からビジョンへのモジュールに成功裏に展開されました。
ビジョンシステムは、3次元（3D）ボクスル表現を生成し、当実験環境において物体姿勢の識別において有効でした（使用したGrounding DINO検出モジュールは、COCOゼロショット転移ベンチマークで平均精度52.5を達成しました）。例えば、モジュールが正しく識別できた例として、

白いカップは、実験条件下で100%使用しました。3Dボックセル表現には、さまざまなオブジェクトのメッシュが含まれていました。これらのメッシュから、1/3 Hzの頻度でターゲットポーズが抽出されました。原則として、システムは任意の物体を検出できるはずだった。しかし、パイロット研究では、ホットドリンクの製造に関連する異なる物体を常に正確に識別できないことが判明した。これは、形状が類似した物体の混同や、トレーニングデータセットに存在しない物体によるものが多かった。また、ロボットのエンドエフェクタによる遮蔽が、物体検出の誤差を引き起こし、混雑した環境で使用した場合にエラーが発生する可能性もあることがわかった。例えば、白色カップの平均成功識別率は、遮蔽率20%から30%の範囲では約90%でしたが、遮蔽率がさらに高くなると大幅に低下し（例えば、遮蔽率80%から90%の範囲では約20%に低下しました）。コンピュータビジョン技術の向上により、ロボットが視覚的に最も複雑な環境にも対応できる能力が向上すると予想される。ただし、ビジョンシステムの性能は印象的で、特定された問題（例えば、分布外オブジェクトの使用）を回避すれば、比較的制約の少ない環境でもシステムが適切に対応できることが確認された（図3）。

Force feedback

ATI多軸力・トルクセンサーが、巧妙な物体相互作用に必要な十分な力フィードバックを提供することを確認した。センサーは6成分の力とトルクを測定し、タスク実行中にロボットのエンドエフェクターが加えた力は成功裏に測定された。センサーの精度はいずれも、サンプリングレート100Hzでフルスケールの約2%以内の誤差範囲内にあった。

タスク実行中に、ロボットは多様な運動ダイナミクスと異なる種類の力フィードバックを伴う動作を示した。図4は、ロボットがコーヒーを準備し、ペンを手渡す際に経験した力を示している。図4に示すように、多様なタスクにおいて多様な外部力が処理されました。例えば、マグカップを置く際には、ピークの上向きの力が成功の指標として使用されました。

一方、引き出しの操作時には、x軸とy軸に沿った力とトルクが重要であり、タスク実行の成功に不可欠であることが示されました。力フィードバックの変動性は、多様な動作の要件に適応するスケーラブルなアプローチの利点を示しています。
注ぎ精度はおおよそ5.4 g/100 g（ピッチ速度4 m s^{-1}）でした。注がれた水の体積を推定するために、準静的平衡を仮定しました。しかし、ピッチ速度が増加するにつれ精度が低下し、ピッチ速度30 m s^{-1}では誤差がおおよそ20 g s^{-1}に近づきました。この精度の低下は、準静的仮定の破綻と、注ぐ媒体と容器の質量分布が測定精度に与える影響に起因すると考えられる。

Generating art

DALL-E^{59}は、画像から描画軌道を導き出すことが可能であることが確認された。これにより、ロボットはユーザーが指定した任意のデザインを描画することが可能になった。DALL-Eは、ユーザーから抽出されたキーワード（例：「ランダムな鳥」や「ランダムな植物」）に基づいてシルエットを生成することができた。シルエットの輪郭は抽出され、ターゲット表面の寸法に合わせるように変換された。これにより、ロボットはさまざまな物理的オブジェクトにデザインを再現することができました（図5）。描画時に力フィードバックが均一なペン圧を適用することが判明し、これによりZ成分の制御が可能になりました（補足セクション2）。

Evaluation

私たちは、RAGや力覚フィードバックを利用しないVoxPoserに対して、ロボット計画生成手法を評価した。手法の比較のため、知識ベースで指定されたタスクの範囲を反映した80の人間のようなクエリをLLMに生成させた。これらのクエリを用いてロボット計画を生成した。RAG（当社の方法）—知識ベースがLLMの意思決定プロセスに動的に統合される—と、知識ベースがLLMのコンテキストウィンドウに静的に組み込まれるベースライン（VoxPoser）の性能結果を比較した。重要な点は、後者のアプローチはスケーラビリティに欠け、知識ベースが拡大するにつれて実践的ではなくなることだ。

私たちは、回答の真実性と正確性を評価する「回答の信頼性」に基づいて結果を評価した（事実の正確な表現を確保し、捏造や「幻覚」エラーを排除する）。私たちの結果では、RAGを使用することで回答の信頼性が向上した。GPT-4（gpt-4-0613）では、RAGを使用することで信頼性スコアが0.74から0.88に増加した。同様に、GPT-3.5-turbo（gpt-3.5-turbo-0125）はRAGを使用した場合0.86、

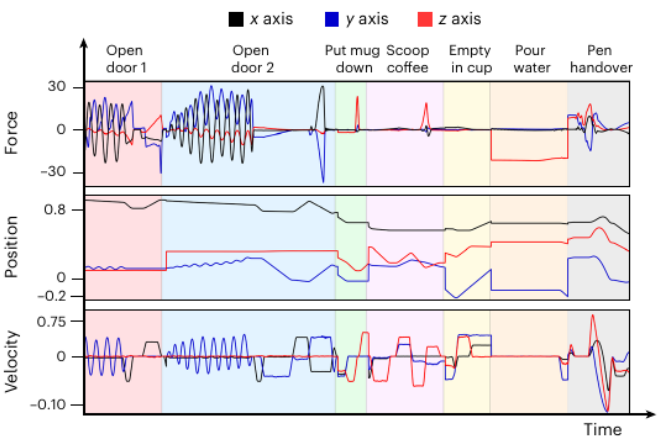

図4 | 力、速度、位置のフィードバック。ロボットのコーヒー準備中の力（N）、速度（m s^{-1}）、位置（m）のグラフ。異なる動作における多様な力フィードバックを示している。明瞭さを重視し、描画コンポーネントは省略した。

https://doi.org/10.1038/s42256-025-01005-x

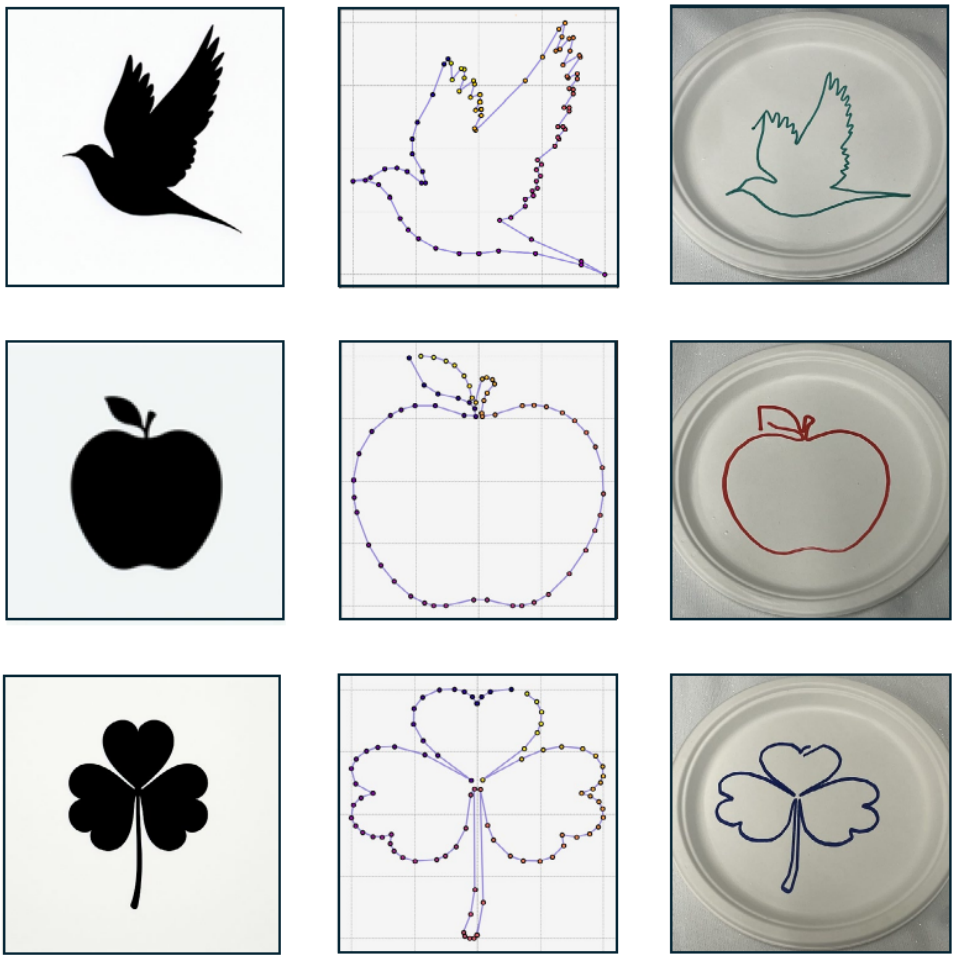

図5 | 描画プロセスの可視化。異なるクエリにおける描画プロセスのイラスト。上段は「ランダムな動物」を作成するように指示された際の生成画像、輪郭プロット、および描画結果を示す。2段目は「ランダムな食べ物」に対する対応する出力、3段目は「ランダムな植物」の結果をイラストで示す。

使用しない場合0.78となり、Zephyr-7B-betaは0.37から0.44に増加しました。忠実度の向上は、物理的な相互作用中の正確な実行が不可欠なロボット応用において特に重要です。

Discussion

私たちは、人工知能とロボット操作の技術を組み合わせたインテリジェントロボットを作成する手法であるELLMERフレームワークをテストした。私たちのアプローチは、LLMsの認知能力とロボットの感覚運動スキルを成功裏に組み合わせ、ロボットが高次言語コマンドを解釈し、複雑な長期タスクを実行しながら不確実性を適切に管理する能力を実現した。LLMにフィードバックループとRAGを組み合わせて、表現力豊かなコードを記述し、ロボットが最終目標（温かい飲料の製造）を達成するために必要な操作サブタスクを実行できるようにした。ELLMERは環境変化へのリアルタイム適応を可能にし、RAGを介して精密な解決策のレポジトリを活用することで、タスクの正確な実行と広範な適応性を確保した^{32}。

ELLMERは既知の制約をコード例（『モーション関数』）にエンコードし、材料の量の変動や未知の引き出しの開閉など、他の手法では広範な追加トレーニングなしでは実現できない多数の不確実性への迅速な適応を可能にしました^{29,33,60,61}。ビジョン、力、言語のモダリティの統合は、操作性能を向上させました。

力センサーはタスクの精度を向上させ（例えば、ビジョンが遮断された際に正確な量の液体を注ぐなど）、ビジョンシステムは物体の位置と動きを識別した。言語機能により、システムはコード内でフィードバックを生成でき、これは新しいタスクへの適応に不可欠だ。キュレーションされた知識ベースは、タスクの仕様に情報検索を最適化することでLLMの性能を向上させ、高品質で文脈に適した出力を確保した。キュレーションされた知識ベースは、制御性、精度、スケーラビリティを向上させる実践的な要素だ。この文脈において、RAGはロボットが知識を引き出すための文化的背景を提供するものと見なせる。特に、これは人間が知識の文化的伝達を通じて獲得する「知能」を反映している。したがって、私たちの研究は、高度な言語モデルとセンサーモータ制御戦略を統合することで、ロボットがLLMの指数関数的進歩を活用し、より高度な相互作用を実現できることを示している。これは、前例のない自律性と精度を備えた自動化の次なる時代を導き、これらの進歩を安全に管理する必要性を強調する^{62}。

ELLMERの潜在能力は、複雑で芸術的な動きの創造にも及ぶ。例えば、DALL-Eのようなモデルは、視覚入力から軌道を導き出し、ロボットの軌道生成に新たな可能性を開く。この方法は、ケーキのデコレーションやラテアートなどのタスクに広く応用可能だ。今後の研究では、クエリと画像の組み込みにより、新たな軌道生成が可能となり、汎用性が向上する。

図6 | コーヒーと皿の装飾。Kinova Gen3ロボットがコーヒーを準備し、皿を装飾した様子。

さらに、最近のLLMの機能向上は、人間とロボットの相互作用の滑らかさと効果を著しく向上させる見込みだ。コーヒーの抽出や皿の装飾の例は、高度なロボットが求められる複雑なタスクの一部に過ぎない。ELLMERはスケーラビリティに優れているため、多様な長期的なタスクに対応可能だ。したがって、ELLMERはフィードバックループのデータベースや「デモンストレーションから学習する」例を組み込むことで、多様な複雑なロボット操作を可能にする可能性がある。

ELLMERは、コンピュータビジョンに関する2つの仮定に基づいています：（1） ビジョンモジュールがシーン内のオブジェクトを正確に識別し分類する （2） 調理器具の包括的なアフォーダンスマップが利用可能である。当モデルには、ケトル、スプーン、ドアノブの アフォーダンスに関する事前知識を付与したが、最近の研究では、アフォーダンスは最小限のデータで学習可能であることが示されている^ {63,64} 。当研究の焦点はオブジェクト検出にはなかったが、検出応答時間が最適性能を妨げる要因となることを確認した。さらに、ELLMERはリアルタイムの変化に適応できましたが、事前プログラミングなしでのタスク切り替えなど、能動的な適応には苦労しました。今後の改良では、言語モデルへの問い合わせ頻度を増加させることで、新しい入力に基づいて全体計画の再評価と修正が可能になります。さらに、複雑な力学の高度なモデリング（例えば、流量、容器のサイズ、液体の粘度に応じてエンドエフェクタに作用する力）や、空間認識ツール（例えば、3D占有マップ用のロボットライブラリであるOctoMaps）の統合など、解決すべき課題が残っている。触覚センサーの組み込みとソフトロボティクス技術の利用は、ロボットが損傷を引き起こさずに適切な力を適用する能力を向上させる。ELLMERは、これらの研究成果を組み込むための柔軟なプラットフォームを提供し、ロボットが「感覚」フィードバックを利用して材料の特性を解釈し、適用する力を精密に調整することを可能にする。

ELLMERの現在のバージョンでは、ロボットが「ワンショット」で複雑なタスクを成功裏に完了することができました。これは、センサーモータスキルとLLMsが提供する抽象的な推論を組み合わせた知能機械の能力を示す説得力のある例です。ただし、ELLMER内で統合されるコンポーネントがさらに精緻化されるにつれ、ロボットの能力が指数関数的に向上すると予想されます。当フレームワークはハードウェアに依存せず、Haystackのようなオープンソースの RAGソリューションで容易にカスタマイズ可能で、エンベッダー、リトリバー、チャンキング技術、LLMの迅速な調整をサポートする。ELLMERは、研究者が協働して知能機械を開発するための柔軟なフレームワークを提供する。補足セクション3では、ELLMERと今後の研究に関する詳細情報を提供する。

私たちのアプローチの力は、強化されたセンサーモータ能力とLLMsの認知推論能力を組み合わせたフレームワークを通じて認知を具現化することにある。この組み合わせにより、ELLMERはロボットが環境を探索し相互作用する能力を向上させ、人間の知能で観察される経験と行動のつながりの一部を模倣する。

これにより、ロボットが「物理的知能」を獲得する可能性が開かれ、環境の探索がセンサーモータ学習プロセスを駆動するようになる。結論として、ELLMERは言語処理、RAG、力学、視覚を統合し、ロボットが複雑なタスクに適応できるようにする。以下の特徴を組み合わせている：（1） 高次の人間コマンドの解釈、（2） 長期的タスクの完了、（3） 変化する環境におけるノイズと干渉を管理するための統合された力学と視覚信号の活用。ELLMERは、強化学習、模倣学習、柔軟な運動プリミティブを統合的に組み合わせることで、多様な動的シナリオにおける適応性と「ロボット知能」を向上させます。LLMの認知推論能力とロボットのセンサーモータスキルを統合することで、環境を解釈・操作し、体現化された機械知能を通じて複雑なタスクを完了できることを示しています。

Methods

Overview

ロボットの目標は、家庭のキッチンなどの動的環境において、高レベルのヒューマンコマンドに応答することでした。私たちは、ケトル、白いマグカップ、引き出し、キッチン用品、コーヒーポットなど、現実的な設定を設計しました。このシナリオは、人間が存在する環境で、現実的ながらも合理的に制約された環境において、ロボットが多様なタスクを実行する能力をテストするために設計された。ロボットの低レベル制御メカニズムが障害物回避を管理すると仮定した。パイプラインは、タスク実行のための言語処理コンポーネント、姿勢検出のためのビジョンシステム、物体操作のための力モジュールから構成されていた。これらすべては、ロボットオペレーティングシステム（ROS）プロセス内に統合されていた。

具体的には、適応可能なロボット動作を可能にする「動的ポリシーのためのコード」アプローチ^ {65} を基盤としています。実装では、GPT-4 と OpenAI の RAG インフラストラクチャを活用しました。RAG^ {32} を使用して LLM の能力を活用し、データベースから最も適切なポリシーを動的に選択・適応させたり、関連する例に基づいて独自のコードを生成したりしています。既存の純粋なLLM駆動型手法^ {25,27,29} と異なり、私たちは力と視覚をフレームワークに統合し、システムが動的環境における多様な複雑なタスクに適応できるようにしました。このアプローチは、ロボットシステムに高レベルの文脈理解能力^ {25} と、リアルタイムフィードバックに基づく複雑なタスクの実行能力を付与し、精度と正確性を確保します。このアプローチにより、各行動はタスクの特定の要件と環境条件に一致するようになります（図6）。

ハードウェアとソフトウェア

Kinovaの7自由度ロボットを使用した。Azure Kinectセンサーを640 × 576 px^ {2} の解像度と30 fpsで、ATI多軸力センサーと組み合わせて使用した。ロボットの先端に140mmのRobotiqグリッパーを装着した。力センサーは、RobotiqグリッパーとKinovaアームに3Dプリントされたフランジを使用して取り付けられた。グリッパーに最も近い側の力センサーに小さなシリンダーを配置し、グリッパーの動作が力センサーに触れないようにした。これにより、測定値の精度が低下するのを防いだ。Intel Core i9プロセッサーとNVIDIA RTX 2080グラフィックプロセッシングユニットを搭載したDellデスクトップコンピュータを使用し、イーサネットケーブルでロボットに接続した。同様に、両方のAzureカメラをデスクトップに固定した。Ubuntu 20.04とROSを使用した。当社のコードはKinova ROS Kortexライブラリに依存している。NVIDIA RTX 2080は、通常の負荷条件下で~225 Wを消費する^ {66} のに対し、Kinovaロボットアームは~36 Wを消費する（参照67）。当社のシナリオでは、各タスクは最大4分間実行されます。EPAの混合エネルギー源に対する平均変換係数（約0.4 kgのCO_ {2} /kWh）^ {68} を用いると、各タスクの二酸化炭素排出量は約0.007 kg（7 g）のCO_ {2} となります。

言語処理

LLMは画像とユーザーのクエリを処理し、複雑なタスクL_ {T} をステップのシーケンス {L_ {1} , L_ {2} , …, L_ {N} } に体系的に分解する。

各ステップL_{i}は、前のステップの完了に依存する場合がある。ステップの順序は重要であり、ステップ間には依存関係がある。例えば、必要な物体（例えばマグカップ）が見つからない場合、キャビネットを開ける必要がある可能性がある。
初期画像入力から収集された環境データは、抽象的なタスクを分解する上で鍵となる。例えば、飲料を作るように指示された場合、環境内に存在する材料はどの飲料を作るかを決定する上で重要であり、視覚情報は可能な位置を特定するのに役立つ。インターフェースはGPT-4により実現され、サーバープラットフォーム経由でロボットにコードを記述して送信する指示の下で動作した。このプロセスは、コード例を含む知識ベースによって支援され、ロボットとの継続的な通信を可能にしました。キュレーションされた知識ベースには、既知の不確実性を組み込んだ低次および高次アクションの検証済み例が含まれていました。これらの動作例を含めることは、ロボットが多数のシナリオに対応し、長期的なタスクを完了するために不可欠です。高次元の動作プリミティブまたはポリシーは、複数の既知の不確実性を単一の関数に圧縮し、広範なコード記述の必要性を削減します。RAGは、パフォーマンスを犠牲にすることなく、知識ベースを包括的に保つことを可能にしました。システムはROSと相互作用し、EC2サーバーが提供する低遅延接続を介してJSONアクションクエリとレスポンスで通信した。

タスク間の依存関係は、P（L_{2A} ， L_{2B} ¤L_{1} ）のような条件付き確率で表現され、タスク L_{1} の成功実行後にタスク L_{2A} または L_{2B} に進む確率を指定します。これにより、ステップのシーケンスを計画し、ロボットがリアルタイムのフィードバックに基づいて動作を適応させることができる。LLMは、指示（プロンプト）と例を含む知識ベースに基づいて実行可能なコードを生成し、サーバーに送信する。コードは、事前定義された関数にのみアクセス可能なセキュアな環境でROS上で実行され、タスクの実行安全性が確保される。

RAG

当社のシステムの重要な特徴の一つは、RAG の展開だ。RAG は、ユーザーからのクエリと継続的に更新されるキュレーション済みの知識ベースからの情報を統合し、LLM の出力を最適化する。このアプローチにより、モデルはデータベースに提供されたコード例に従うことができ、知識ベースが進化しても精度、信頼性、スケーラビリティを確保できる。
私たちはベクトルRAGを使用しました。これは、クエリ（q）と知識ベースのセグメント（ {s_{1} ， s_{2} ，…， s_{m} } ）をチャンクとしてベクトル表現に埋め込むエンコーダーを使用する手法です。チャンクはコサイン類似度に基づいてクエリと比較され、上位k個のチャンクが文脈的に関連する情報として選択され、応答を生成するために使用されます。私たちのフレームワーク内で使用できる代替の検索技術には、従来のRAG（キーワード/ルールベースのRAG）やハイブリッド検索手法がある。

RAGパイプラインは、異なるドキュメントストア（知識ベースが格納され組織化されるメディア）を選択することでカスタマイズ可能です。当社の実験テストでは、組み込みのOpenAI RAGプロセスを使用し、キュレーションされた知識ベースをマークダウンファイルとしてドキュメントストアに組織化しました。ただし、当フレームワークでは、Haystack^{56} や Vebra^{57} などのツールを活用し、他の多様な RAG アプローチも利用可能です。これらのツールは、ユーザーがドキュメントストア（シンプルなテキストベースの知識用の「マークダウンファイル」から、複雑なインデックス付きデータ用の「Elasticsearch」まで）に加え、特定のエンベッダー、リトリバー、チャンキング技術、および LLM 自体を選択できるようにします。

Vision system

Grounded-Segment-Anythingを言語からビジョンへのモデルとして使用し、すべてのオブジェクトの位置を強調表示し、ロボットの把持のためにその姿勢を抽出可能な3Dボクスセルを生成した^{58,69}。これにより、（1） オブジェクト固有のバウンディングボックスの生成、（2） MobileSAMによるセグメント化されたマスクの生成、（3） 検出されたオブジェクトをカプセル化するボクスセルの生成が可能になった。ボクセルにより、ターゲットオブジェクトの姿勢を抽出することが可能になりました。

Force module

力豊かなアプリケーションでの正確な測定を確保するため、ATI力センサーを重力の影響を補償するように校正し、外部力が作用しない状態でゼロを示すようにした。この校正は、エンドエフェクタに作用する外部力を正確に予測するために不可欠だ。プロセスは、力センサーを1軸でゼロ調整し、センサーを回転させて次にゼロ調整する手順を順次繰り返すものだった。局所的な力はグローバル平面に変換され、異なる回転角度における上向きの力を推定するために使用されます。F_{global} = T_{end_effector_to_robot_base} × F_{local}、ここで F_{global} はロボットベース座標系における力ベクトル、T_{end_effector_to_robot_base} はエンドエフェクタの座標系からロボットのベース座標系への変換行列、F_{local} はエンドエフェクタのローカル座標系における力ベクトルです。センサーの位置や方向を移動させる方法や、多項式関数を用いた校正方法など、さまざまな方法を検討しました。しかし、最も効果的だったのは、よりシンプルな校正方法でした。
流量を推定するため、静的平衡状態を仮定し、注ぎ込み中は低速で操作を継続する。数学的には、これはF_{up} ≈ mgとΔF_{up} ≈ Δmgで表される。変数加速度を含む状況では、力と流量の関係はより複雑になる。流量、容器の重心、エンドエフェクタの慣性などの変動する入力を考慮した動的モデルが必要となり、動的力入力を注ぎ流量にマッピングする必要がある。

システムは、3つの軸に沿って力ベクトルを継続的に管理し、知識ベース内の基準に基づいて適用する力を調整する。LLMは、特定のダウンストリームタスク要件を満たすために、必要な力の方向と大きさを動的に選択する。例えば、知識ベースは、対象物の特性やタスクの要件に応じて適用する力の大きさを指定することができる。このアプローチにより、システムは幅広い運用基準に自律的に適応して行動を調整することができる。

ROS operation

本研究では、Kinova ROS Kortexドライバーを起動してロボットプロセスを開始した。これにより、ROSネットワークとKinova Gen3ロボット間の通信を可能にするノードが確立された。このノードは、サブスクライバーがアクセス可能な複数のトピックをパブリッシュし、ロボットの構成を変更するためのサービスを呼び出せるように提供する。ベースジョイントは40Hzの頻度で更新される。同時に、Robotiq 2F-140 mmグリッパーノードは50Hzでアクティブ化される。ノードはUSB接続を介してグリッパーとの通信リンクを設定し、グリッパーの精密制御を可能にし、動作データの交換を容易にするアクションサーバーを起動する。

私たちのロボットシステムの重要な要素は、ビジョンモジュールノードです。環境内の選択されたオブジェクトのターゲット姿勢を特定するために、『classes』変数が使用されます。この変数は動的に更新可能であり、これによりシステムはシーンの変化に適応できる。『classes』変数で確立されたオブジェクトの姿勢座標は、約_{3}^{1} Hzの頻度で公開される。これは主に、Grounding DINOがオブジェクトを検出およびバウンディングボックスを確立する処理時間に起因する。さらに、カメラのロボットベースに対する位置を決定するためにAprilTagを使用した。これはP^{R} = T_{AR} × （T_{CA} × P^{C} ）で表される。ここで、P^{C}はカメラフレーム内の点、T_{CA}はカメラフレームからAprilTagへの変換行列、T_{AR}はAprilTagからロボットのベースへの変換行列、PRはロボットのベースフレーム内の点である。
並行して、100 Hzの周波数で力ノードが起動し、ATI力変換器に局在化された多軸の力とトルクの読み取り値を提供する。測定値は、ロボットのグローバルベースフレームに一致させるため、クォータニオンベースの3×3回転行列を使用して変換され、固定自由度における直近5つの時間ステップの raw 値と平均値が提供される。ロボットベースのグローバルフレームにおける力を、運動学データから計算された回転行列を使用して計算する。

ROSは、言語処理、ビジョンシステム、力メトリクス、および関節エンドエフェクタの位置から得られるマルチモーダルフィードバックデータの連続処理を可能にする。動作は、速度と可変速度および力グリップ手順（開く/閉じる）を制御する基礎的な6自由度ツイストコマンドに基づいて実行される。これにより、最大速度や力制限などのハードコーディングされた安全制約や作業領域境界の統合が実現される。

線形速度は±0.05 m s^{-1}以内に、角速度は±60° s^{-1}以内に制限された。エンドエフェクタの力は20 Nに制限された。これは基本動作プリミティブにコード化されているため、言語モデルの誤りがこれを上書きすることはない。エンドエフェクタは、事前定義された作業領域の境界 x = [0.0, 1.1], y = [-0.3, 0.3], z = [0, 1.0] 内に固定されている。これは、10 Hz の頻度でパブリッシャによって今後の時間ステップでチェックされる。

Data availability

本研究で使用したデータセットは、オープンソースの GitHub リポジトリ（https://github.com/ruaridhmon/ELLMER）から入手可能です。

Code availability

The code supporting this study is available via GitHub at https://github.com/ruaridhmon/ELLMER and has been archived in Zenodo at https://doi.org/10.5281/zenodo.14483539 (ref. 70).

References

1. Intelligence research should not be held back by its past. *Nature* **545**, 385–386 (2017).
2. Friston, K. Embodied inference and spatial cognition. *Cogn. Process.* **13**, 497–514 (2012).
3. Wilson, M. Six views of embodied cognition. *Psychon. Bull. Rev.* **9**, 625–636 (2002).
4. Clark, A. An embodied cognitive science. *Trends Cogn. Sci.* **3**, 345–351 (1999).
5. Stella, F., Della Santina, C. & Hughes, J. How can LLMs transform the robotic design process? *Nat. Mach. Intell.* **5**, 561–564 (2023).
6. Miriyev, A. & Kovac, M. Skills for physical artificial intelligence. *Nat. Mach. Intell.* **2**, 658–660 (2020).
7. Cui, J. & Trinkle, J. Toward next-generation learned robot manipulation. *Sci. Robot.* **6**, eabd9461 (2021).
8. Arents, J. & Greitans, M. Smart industrial robot control trends, challenges and opportunities within manufacturing. *Appl. Sci.* **12**, 937 (2022).
9. Billard, A. & Kragic, D. Trends and challenges in robot manipulation. *Science* **364**, eaat8414 (2019).
10. Yang, G.-Z. et al. The grand challenges of *Science Robotics*. *Sci. Robot.* **3**, eaar7650 (2018).
11. Buchanan, R., Rofer, A., Moura, J., Valada, A. & Vijaya kumar, S. 視覚と固有感覚センシングを用いた因子グラフによる関節構造物体のオンライン推定. 2024 IEEE International Conference on Robotics and Automation (ICRA) 16111-16117 (IEEE, 2024).
12. Nikolaidis, S., Ramakrishnan, R., Gu, K. & Shah, J. 人間-ロボット協働タスクのための関節動作デモからの効率的なモデル学習. 2015年 第10回 ACM/IEEE 人間-ロボット相互作用国際会議 (HRI) 189-196 (IEEE, 2015).
13. Saveriano, M., Abu-Dakka, F. J., Kramberger, A. & Peternel, L. Dynamic movement primitives in robotics: a tutorial survey. Int. J. Robot. Res. 42, 1133-1184 (2023).
14. Kober, J. et al. Movement templates for learning of hitting and batting. In *2010 IEEE International Conference on Robotics and Automation* 853–858 (IEEE, 2010).
15. Huang, W. et al. VoxPoser: composable 3D value maps for robotic manipulation with language models. In *Proc. 7th Conference on Robot Learning* 540–562 (PMLR, 2023).
16. Zhang, D. et al. Explainable hierarchical imitation learning for robotic drink pouring. In IEEE Transactions on Automation Science and Engineering 3871-3887 (2022).
17. Hussein, A., Gaber, M. M., Elyan, E. & Jayne, C. Imitation learning: a survey of learning methods. *ACM Comput. Surv.* **50**, 21:1–21:35 (2017).
18. Di Palo, N. & Johns, E. DINOBot: robot manipulation via retrieval and alignment with vision foundation models. In *International Conference on Robotics and Automation (ICRA)* 2798–805 (IEEE, 2024).
19. Shridhar, M., Manuelli, L. & Fox, D. CLIPort: what and where pathways for robotic manipulation. In *Proc. 5th Conference on Robot Learning* 894–906 (PMLR, 2022).
20. Shridhar, M., Manuelli, L. & Fox, D. Perceiver-Actor: a multi-task transformer for robotic manipulation. In *Proc. 6th Conference on Robot Learning* 785–799 (PMLR, 2023).
21. Mees, O., Hermann, L. & Burgard, W. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robot. Autom. Lett.* **7**, 11205–11212 (2022).
22. Mees, O., Borja-Diaz, J. & Burgard, W. Grounding language with visual affordances over unstructured data. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* 11576–11582 (IEEE, 2023).
23. Shao, L., Migimatsu, T., Zhang, Q., Yang, K. & Bohg, J. Concept2Robot: learning manipulation concepts from instructions and human demonstrations. *Int. J. Robot. Res.* **40**, 1419–1434 (2021).
24. Ichter, B. et al. Do as I can, not as I say: grounding language in robotic affordances. In *Proc. 6th Conference on Robot Learning* 287–318 (PMLR, 2023).
25. Driess, D. et al. PaLM-E: an embodied multimodal language model. In *Proc. 40th International Conference on Machine Learning* 8469–8488 (PMLR, 2023).
26. Peng, A. et al. Preference-conditioned language-guided abstraction. In *Proc. 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI '24* 572–581 (Association for Computing Machinery, 2024).
27. Huang, W., Abbeel, P., Pathak, D. & Mordatch, I. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In *Proc. 39th International Conference on Machine Learning* 9118–9147 (PMLR, 2022).
28. Huang, J. & Chang, K. C.-C. Towards reasoning in large language models: a survey. In *Findings of the Association for Computational Linguistics: ACL 2023* 1049–1065 (Association for Computational Linguistics, 2023).
29. Zitkovich, B. et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In *Proc. 7th Conference on Robot Learning* 2165–2183 (PMLR, 2023).
30. Ma, X., Patidar, S., Haughton, I. & James, S. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18081–18090 (IEEE, 2024).
31. Zhang, C., Chen, J., Li, J., Peng, Y. & Mao, Z. Large language models for human-robot interaction: a review. *Biomimetic Intell. Robot.* **3**, 100131 (2023).
32. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* 9459–9474 (Curran Associates, 2020).
33. Raiaan, M. et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access* **12**, 26839–26874 (2024).
34. Rozo, L., Jimenez, P. & Torras, C. Force-based robot learning of pouring skills using parametric hidden Markov models. In *9th International Workshop on Robot Motion and Control* 227–232 (IEEE, 2013).

35. Huang, Y., Wilches, J. & Sun, Y. Robot gaining accurate pouring skills through self-supervised learning and generalization. *Robot. Auton. Syst.* **136**, 103692 (2021).
36. Mon-Williams, R., Stouraitis, T. & Vijayakumar, S. A behavioural transformer for effective collaboration between a robot and a non-stationary human. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* 1150–1157 (IEEE, 2023).
37. Belkhale, S., Cui, Y. & Sadigh, D. Data quality in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)* 80375–80395 (Curran Associates, 2024).
38. Khazatsky, A. et al. DROID: a large-scale in-the-wild robot manipulation dataset. Robotics: Science and Systems; https://www.roboticsproceedings.org/rss20/p120.pdf (2024).
39. Acosta, B., Yang, W. & Posa, M. Validating robotics simulators on real-world impacts. *IEEE Robot. Autom. Lett.* **7**, 6471–6478 (2022).
40. Alomar, A. et al. CausalSim: a causal framework for unbiased trace-driven simulation. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* 1115–1147 (USENIX Association, 2023).
41. Choi, H. et al. ロボット工学におけるシミュレーションの利用について：機会、課題、および今後の提案。Proc. Natl Acad. Sci. USA 118, e190785611 (2021).
42. Del Aguila Ferrandis, J., Moura, J. & Vijayakumar, S. Nonprehensile planar manipulation through reinforcement learning with multimodal categorical exploration. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 5606–5613 (IEEE, 2023).
43. Kirk, R., Zhang, A., Grefenstette, E. & Rocktäschel, T. ディープ強化学習におけるゼロショット一般化に関する調査. J. Artific. Intell. Res. 76, 201-264 (2023).
44. Dai, T. et al. ドメインランダム化で訓練された深層強化学習エージェントの分析. Neurocomputing 493, 143-165 (2022).
45. Chang, J., Uehara, M., Sreenivas, D., Kidambi, R. & Sun, W. 部分的なカバレッジを持つオフラインデータによる模倣学習における共変量シフトの軽減。Advances in Neural Information Processing Systems 965-979 (Curran Associates, 2021).
46. Huang, W. et al. 内なる独白：言語モデルを用いた計画による身体化された推論. 第6回ロボット学習会議プロシーディングス 1769-1782 (PMLR, 2023).
47. Nair, S., Rajeswaran, A., Kumar, V., Finn, C. & Gupta, A. R3M: a universal visual representation for robot manipulation. In *Proc. 6th Conference on Robot Learning* Vol. 205, 892–909 (PMLR, 2022).
48. Singh, I. et al. ProgPrompt: generating situated robot task plans using large language models. In *Proc. IEEE/CVF International Conference on Robotics and Automation (ICRA)* 11523–11530 (IEEE, 2023).
49. Song, C. H. et al. LLM-Planner: few-shot grounded planning for embodied agents with large language models. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 2998–3009 (IEEE/CVF, 2023).
50. Vemprala, S. H., Bonatti, R., Bucker, A. & Kapoor, A. ChatGPT for robotics: design principles and model abilities. *IEEE Access* **12**, 55682–55696 (2024).
51. Ding, Y., Zhang, X., Paxton, C. & Zhang, S. 大規模言語モデルを用いた物体の再配置のためのタスクと動作計画. 2023 IEEE/RSJ 国際知能ロボットシステム会議 (IROS) 2086-2092 (IEEE, 2023).
52. Kwon, M. et al. 現実に基づく常識的推論に向けて。Proc. International Conference on Robotics and Automation (ICRA) 5463-5470 (IEEE, 2024).
53. Hong, J., Levine, S. & Dragan, A. Learning to influence human behavior with offline reinforcement learning. In Advances in Neural Information Processing Systems (NeurIPS) 36094-36105 (Curran Associates, 2024).
54. OpenAI. GPT-4 technical report. Preprint at http://arxiv.org/abs/2303.08774 (2024).
55. OpenAI. Custom models program: fine-tuning GPT-4 for specific domains (2023); https://platform.openai.com/docs/guides/fine-tuning/
56. Pietsch, M. et al. Haystack: the end-to-end nlp framework for pragmatic builders. *GitHub* https://github.com/deepset-ai/haystack (2019).
57. Weaviate. Verba: the golden RAGtriever. *GitHub* https://github.com/weaviate/Verba (2023).
58. Kirillov, A. et al. Segment anything. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 4015–4026 (IEEE, 2023).
59. Ramesh, A. et al. Zero-shot text-to-image generation. In *Proc. 38th International Conference on Machine Learning* 8821–8831 (PMLR, 2021).
60. Zeng, A. et al. Socratic models: composing zero-shot multimodal reasoning with language. In *Proc. International Conference on Learning Representations* (ICLR, 2023).
61. Cui, Y. et al. No, to the right: online language corrections for robotic manipulation via shared autonomy. In *Proc. 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI '23* 93–101 (Association for Computing Machinery, 2023).
62. Bengio, Y. et al. Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845 (2024).
63. Li, G., Jampani, V., Sun, D. & Sevilla-Lara, L. Locate: localize and transfer object parts for weakly supervised affordance grounding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10922–10931 (IEEE, 2023).
64. Li, G., Sun, D., Sevilla-Lara, L. & Jampani, V. One-shot open affordance learning with foundation models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3086–3096 (IEEE, 2024).
65. Liang, J. et al. Code as policies: language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* 9493–9500 (IEEE, 2023).
66. Hong, S. & Kim, H. An integrated GPU power and performance model. In *Proc. 37th Annual International Symposium on Computer Architecture* 280–289 (Association for Computing Machinery, 2010).
67. Kinova Robotics. Kinova Gen3 Ultra-Lightweight Robotic Arm User Guide (2023); https://assets.iqr-robot.com/wp-content/uploads/2023/08/20230814163651088831.pdf
68. US Environmental Protection Agency. GHG emission factors hub (2024); https://www.epa.gov/climateleadership/ghg-emission-factors-hub
69. Liu, S. et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *2024 European Conference on Computer Vision* (eds Leonardis, A. et al.) Vol. 15105 (Springer, 2023).
70. ruaridhmon. ruaridhmon/ELLMER: v1.0.0: Initial Release. *Zenodo* https://doi.org/10.5281/zenodo.14483539 (2024).

Acknowledgements

この研究は、EPSRC CDT in RAS (EP/L016834/1) の支援を受けて実施されました。S. Vijayakumar 氏の支援とリソースへのアクセス提供、L. Martins 氏および Edinburgh Workshop のハードウェアに関する支援、J. Wang 氏、T. Stouraitis 氏、J. Ferrandis 氏、その他多くの方々の貴重な支援と専門知識に感謝いたします。

Author contributions

概念化：R.M.-W., G.L. および R.L. 方法論：R.M.-W., G.L., R.L., W.D. および C.G.L. ソフトウェア：R.M.-W., G.L., R.L. および W.D. 形式分析：R.M.-W. 調査：R.M.-W. および W.D. 可視化：R.M.-W. 検証：R.M.-W. および W.D. 原稿執筆：

R.M.-W. 執筆—レビューと編集：R.M.-W.、G.L.、R.L.、およびC.G.L. 監督：C.G.L.

Competing interests
著者は、利益相反はないことを宣言する。

追加情報
補足情報オンライン版には、https://doi.org/10.1038/s42256-025-01005-x で閲覧可能な補足資料が含まれている。

問い合わせや資料の請求は、Ruaridh Mon-Williams または Gen Li までお願いします。

ピアレビュー情報 Nature Machine Intelligence は、この論文のピアレビューに貢献した Matteo Saveriano およびその他の匿名レビューアに感謝する。