ChatGPT

大規模言語モデルと「ワールドモデル」：最新動向と今後の展望

1. LLMにおける世界シミュレーション能力の技術的実装

大規模言語モデル（LLM）は、**内部に「世界」をシミュレートする能力**を持つとしばしば言われます。技術的には、LLMは**トランスフォーマー**に代表されるニューラルネットワークアーキテクチャ上に構築されており、**自己回帰型**（auto-regressive）の**次トークン予測**によってテキストを生成します [1]。モデルは与えられた文脈から最も尤もらしい次の単語（トークン）を逐次予測していくことで文章を生成します。このメカニズム自体は単純な**パターン補完**に見えますが、膨大なテキストコーパスで事前学習されたLLMの**内部表現**には、驚くほど豊かな世界知識や推論能力が**暗黙的に組み込まれている**ことがわかってきました。

例えば、**事前学習済み言語モデルは知識ベースとして機能しうる**ことが知られており、モデル内部には人名と職業の対応や首都と国といった**関係知識がニューロンレベルで符号化**されています [2]。実際、BERTなどのモデルでは「アゼルバイジャンの首都はバクー」といった事実に対応する特定の**「知識ニューロン」**が存在し、そのニューロンを操作することでモデルの出力する事実を改変できることが報告されています [2]。このようにLLMは**内部に世界の事実関係を分散表現**として学習しており、一種の**記号的な知識ベース**を内包していると捉えられます。

さらに、LLMは文脈内で**動的な状態をシミュレート**することも可能です。興味深い事例として、**Othello（リバーシ）**の棋譜を学習した言語モデル（Othello-GPT）の研究があります [3]。このモデルは次の合法手を予測するよう訓練されただけにもかかわらず、内部に**盤面の状態（石の配置）を表現するベクトル表現が自発的に形成**されたことが確認されました [4] [5]。研究者はモデルの中間層のベクトルから盤面を**線形変換**で読み出すことに成功し、モデルが明示的に与えられなかった**「ゲームの世界モデル」**を内部に獲得していたと結論付けています。この結果は「単なる次手予測タスクでも、モデルはデータを生成する**潜在プロセス（ゲームのルールと盤面変化）を内部に再現している**」ことを示唆するもので、**LLMの内部における世界シミュレーション**の一例と言えます [6] [7]。

では、LLMはどのようにして**テキスト世界の一貫したシミュレーション**を実現しているのでしょうか。鍵はトランスフォーマーの**自己注意機構**（Self-Attention）にあります。自己注意層は入力文脈中の関連する単語同士の関係に重み付けを行い、各単語の**コンテキストに応じた意味表現**を動的に計算します [8] [9]。これによりモデルは文章内の**長距離の関係性**（例えば指示語とその指す対象など）を適切に把握し、文脈に矛盾しない出力を生成できます。また多層のトランスフォーマーブロックによって、**低次の語句レベルの特徴から高次の抽象概念レベルの特徴まで**段階的に内部表現が洗練されます。実際、ある研究では**人間の脳が音声言語を処理する際の階層的な神経活動パターン**が、GPT-2やLLaMA2などLLMの層ごとのテキスト処理過程と時間的・階層的に対応していることが示されました [10]。これは、**モデル内部で言語の意味構造が階層的に表現され蓄積されている**可能性を示唆しています。

以上のように、LLMの世界シミュレーション能力は、**巨大なパラメータ（重み行列）に埋め込まれた知識**と**動的な注意機構による文脈のトラッキング**によって実現されています。その結果、モデルは文章内に明示されていない情報であっても、過去の学習から得た**知識やパターンを総動員して補完**し、人間にとってもっともらしい世界を文章上に再構築します。もっとも、このシミュレーションはあくまでテキスト上でのものであり、**完全に安定した論理的推論や物理的整合性を保証するものではありません**。実際、LLMは複雑な計算問題や厳密な論理推論を内部で逐次解いているわけではなく、あくまで**深く訓練された言語パターンの反射**に基づい

て応答している側面があります [11]。こうした限界があるにもかかわらず、テキスト中の文脈やパターンから**驚くほど汎用的な知識と推論の振る舞いが創発**している点に、現代のLLMの凄みがあります。

2. テキストベースLLMへの批判的見解とその評価（ルカ氏の見解）

LLMの驚異的な能力が明らかになる一方で、**「テキストだけで訓練されたLLMでは脆弱であり行き詰まりだ」**という批判的な意見も専門家から提起されています [12] [13]。例えばMetaのチーフAI科学者である**Yann LeCun（ヤン・ルカン）氏**は、現在主流のLLMアプローチについて「人間レベルの知能に至る道としては行き止まりである」と明言し、この方向に業界が傾きすぎていることへの危機感を示しています [14]。彼は「現在のLLMパラダイムが抱える限界」を具体的に次の4点に整理しています： [15]

- **物理世界の理解の欠如**：純粋にテキストだけを読み込んだモデルは、現実の物理法則や空間的直観を持たないため、常識外れな誤答（例えば「人が空を飛ぶ」ような誤った物理的記述）を平気で生成してしまいます [16]。言い換えれば、**空間や物理的制約に関する「常識」**が身についていないという指摘です。

- **持続的なメモリの欠如**：LLMは一度訓練が終わると内部パラメータが凍結され、**新たな知識をオンラインで学習することができません** [17]。会話中の一時的なコンテキスト（数千トークン程度）は保持できますが、人間のように**経験を蓄積して成長する**ことができないため、長期的な一貫性や学習能力に限界があります。また作業記憶的な**ワーキングメモリ**もコンテキストウィンドウの範囲に限定され、長い文章の前半で提示された情報を後半で忘れてしまうこともあります。

- **推論・論理能力の弱さ**：大規模テキストコーパスからパターンを統計的に学習したLLMは、一見論理的な応答もできますが、複雑な論理推論や数理的な問題では**不安定な振る舞い**を見せがちです [18]。一貫した逐次推論を行う内部メカニズム（例えば人間が紙に書いて推論するような仕組み）が無いため、**ステップバイステップの厳密な推論には脆弱**だという批判があります [19]。この脆弱性は**「幻覚（ハルシネーション）」**と呼ばれる事実無根の回答を自信満々に生成してしまう現象にも現れています [20]。

- **複雑な計画能力の欠如**：現行のLLMは**与えられた入力に対する反応的なテキスト生成**は得意ですが、長期的な目標を立てて行動（出力内容）を計画・実行する能力がありません [15]。例えば、具体的なタスクを完遂するためにサブゴールを設定し中間結果を検証しながら進める、といった**プランニング**は苦手です。これはモデル内部に**目的指向のエージェント機構**や**試行錯誤のフィードバックループ**が無いことに起因します。

以上のような**技術的限界**に加え、哲学的な観点からは「LLMは言語の形式的パターンを真似ているに過ぎず、意味の理解には至っていない」という批判もあります。エミリー・ベンダー氏らの論文になぞらえて**「巨大な確率的オウム」**と揶揄されるように、LLMは膨大なテキストから統計的相関を引き出しているに過ぎず、実世界の意味や因果を理解していないという主張です [8]。実際、ある研究者はLLMの状態を「暗闇の中で数字の羅列だけを見ている存在」に喩え、例えば「怒り」という感情に対応するトークン列を統計的に学習することはできても、**怒りそのものを体験・理解しているわけではない**と指摘しています [21]。このように、**シンボル（記号）操作としての言語処理と、その背後にある意味理解の断絶**がLLMにはあるのではないか、というのが哲学的批判の核心です。

こうした批判に対しては、**反論や別の視点**も存在します。一つは実用面からの評価で、批判派であっても「LLMはAGI（汎用人工知能）には到達しないかもしれないが**実用上の価値は極めて高い**」と認める声です [22]。実際、現在のGPTやClaudeなどのLLMは、プログラミング支援、文書要約、創造的文章生成など、特定タスクにおいて**人間の生産性を飛躍的に高める**ツールとなっています [23]。ある開発者はLLMの有用性を「車のセールスマン」に例えています [24]。セールスマンは自動車の詳細を知り尽くしていても自分で車を造

れるわけではないが、その知識には大いに価値があるように、LLMも**世界を生み出す能力（AGI達成）はなくとも、人間をサポートする有益な知識スキルを提供できる**というわけです。

もう一つの反論は、**LLMの限界を拡張する工夫**が進んでいる点です。批判派が指摘するような欠点（知覚の欠如・記憶の欠如・推論の弱さ）は、研究者たちがまさに認識して取り組んでいる課題でもあります。例えば、**マルチモーダル化**によってテキスト以外の情報を取り入れる試みはその一つです。GPT-4は画像も入力できるようになり、視覚情報を加味した応答が可能になりました。また、外部ツールや知識ベースを参照する**Retrieval-Augmented Generation**、対話中に中間思考を明示させる**チェイン・オブ・ソート（CoT）**プロンプト、セルフリファレンスによる自己検証など、**プロンプト設計や推論支援**の手法によってLLMの弱点を補う研究も数多く登場しています [25]。こうした延長線上では、LLM自体に内蔵された推論モジュールや長期記憶モジュールを組み込む新アーキテクチャ（次節で述べるJEPAやLCMのようなもの）も提案されており、**「テキストLLM＝行き止まり」ではなく進化途中の一形態**だと見る専門家もいます。

実際、Reddit上でLeCun氏の発言を受けた議論でも、「LLM自体には役割がある。ただしAGIに至るものではない」という**穏当な評価**が支持を集めています [22]。総じて、批判派の主張する欠点は事実として受け止めつつ、それを認めた上で**「では次に何をすべきか」**という建設的議論へと今まさに移行しつつある、というのが現状と言えるでしょう [26]。

3. JEPAや次世代アーキテクチャ：「真のワールドモデル」へのアプローチ

批判派の指摘するLLMの限界を克服し、**「真に世界を理解し予測できるAI」**を目指すべく、近年登場しているのが**JEPA（Joint Embedding Predictive Architecture）**をはじめとする新世代のアーキテクチャです。**MetaのYann LeCun氏**が提唱するJEPAは、従来の生成モデルとは一線を画し、**観測データから世界の内部状態を学習・予測する**ことに重きを置いたアプローチです [27]。従来のLLMはテキストの次単語をそのまま生成（予測）する**自己回帰的生成**でしたが、JEPAでは**生データを直接生成せず**に、まずその抽象的な**「表現（エンベディング）」**を捉え、未来や欠損部分の表現を予測するというプロセスを取ります [28] [29]。

図：JEPAに基づくモデルは、物理世界の振る舞いを学習し、常識に反する事象（例：リンゴが上向きに落ちる）に対して「驚き」の反応を示す [30]。このイラストは、物理法則に反する状況を目にしたAI（ニュートンの姿をしたロボット）が、落下するはずのリンゴが逆に空へ昇っていく様子に驚きを感じるコンセプトを表現している。

具体的な例として、**Meta社が開発したV-JEPA（Video JEPA）**を見てみましょう。V-JEPAは動画データから**物理世界のパターン**を学習するモデルで、訓練時に動画の一部領域を隠し（マスクし）、そこから抽象的特徴（**潜在表現**）を抽出するようエンコーダ①に与えます。同時に隠さずに与えた映像から別のエンコーダ②が潜在表現を抽出します。次に**予測器**がエンコーダ①の出力からエンコーダ②の出力を予測するよう学習します [29]。要するに、「一部欠損した観測」から「完全な観測」の**高次元特徴を当てる**訓練をすることで、モデルはマスク部分に何が起きているかを**内部で予測再現できる抽象状態**を獲得します [31] [32]。この方法の利点は、従来のピクセル単位での完全再現を目指す生成とは異なり、**細部の不要な情報を捨てて本質的な構造のみを捉える**点にあります [33] [28]。例えば従来のピクセル予測では、木々の葉の揺れなど重要度の低い要素まで再現しようとしてしまいますが、V-JEPAでは**「車がどこにいて信号機が何色か」**といった重要事項に集中し、枝葉末節は無視する方向に自動で誘導されます [33] [34]。Metaの研究者Quentin Garrido氏は「V-JEPAは**不要な情報を捨てて重要な事象にフォーカスする**ことを効率的に実現する」と述べています [35]。

このアプローチの成果は、興味深い**「驚き」の検出実験**によって示されました。V-JEPAに日常的な物理シーンの動画を見せ、途中で物体があり得ない動きをする映像（例えば、冒頭のイラストのように板がコップをすり抜けるなど）を入力すると、モデルは**通常とは異なる内部予測誤差を示し**、まるで人間の赤ん坊が不可能な事象に驚くような反応を示したのです [30] [36]。これは、モデルが**オブジェクトの永続性**や**物体同士の非**

貫通性といった基本的な物理法則を自らの内部表現に獲得していたことを意味します。実際、生後6ヶ月の乳児は目の前の物体が突然消えたり貫通したりすると驚くことが知られていますが [30]、V-JEPAは**ビデオから類似の物理直観を学習**し、不合理なシーンに対して**予測矛盾（＝驚き）**を示したというわけです [37] [38]。

JEPA系アーキテクチャの狙いは、以上のようにして**世界のメンタルモデル**をAIに獲得させることです。テキストLLMが明示的に持たなかった**「物理的常識」**や**「連続学習能力」**を、観測データの予測というタスクを通じてモデル内に宿らせようという発想です [39] [40]。重要なのは、JEPAでは**予測の対象が元データそのものではなく抽象表現**であるため、モデルは**「この状況で本質的に何が起きているか」**を学習することに集中できます [29] [32]。その結果、JEPAで学習した表現は人間にとって解釈しやすく、たとえば「シーンに何台の車がいて、それらがどの方向に動いているか」といった情報を明示的に含む可能性があります。実際、前述のV-JEPAの研究でも、モデルの潜在ベクトルに物体や動きに関する意味的な次元が現れていることが示唆されています。

JEPA以外にも、**次世代アーキテクチャの例**として注目されるのが**Large Concept Model（LCM）**です。LCMはMeta社を中心に研究が進む新パラダイムで、**単語列としての言語ではなく、文全体や概念単位でのモデリング**を志向しています [41] [42]。簡潔に言えば、LLMが「次の単語」を予測するのに対し、LCMは「次の**概念**や**意味単位**」を操作対象にすることで、より構造的な推論を可能にしようというものです [43]。具体的には、人間の専門家が使うような**知識グラフ**や**因果関係のネットワーク**を内部に取り込み、推論の道筋を明示的に追跡・説明できるAIを目指しています [43] [44]。このアプローチにより、LLMで問題となっている**誤情報や幻覚の混入**を低減し、**信頼性と透明性の高い推論**が可能になると期待されています [43] [45]。InfoQの解説によれば、LCMは単なる言語モデルから**構造化知識に基づく推論エンジン**へとシフトするもので、**因果関係の理解**や**推論過程の可視化**によってAIの判断をより信頼できるものにするという狙いがあります [46] [47]。言い換えれば、LCMは**「知識と推論の融合」**であり、これをLLMの長所（流暢な言語生成能力）と組み合わせることで「高度なシナリオ分析や意思決定を行い説明もできるAI」に近づけるという展望が語られています [48]。

これらJEPAやLCMに共通するのは、**「テキスト予測を超えて世界そのものをモデル化する」**という志向です。前節までに述べた批判の核心は「現在のLLMには世界そのものがない（=ただ言葉の関連性だけを見ている）」という点でした [16]。JEPAやLCMはまさにその点を踏まえ、**視覚・動画など多様なモーダルからの学習** [49]、**内部メモリや知識ベースの組み込み**、**構造的な推論モジュールの統合**といった手段で、次世代の**「真のワールドモデル」**を構築しようとしています [41] [50]。

もちろん、これらのアプローチはまだ発展途上であり、現時点では限定的なデモンストレーション段階です。例えば、V-JEPAの物理直観は動画内の単純な接触や遮蔽に関するもので、人間のような幅広い常識には遠いでしょう。しかし**最初の一歩**として、AIが世界のルールを内部に構築しつつある兆しを示した意義は大きいと言えます。またLCMに関しても、知識グラフを用いた推論AI自体は過去にも試みがありますが、それを大規模モデルスケール・大量知識で実現することで新たな性能ブレークスルーが期待されています [46]。JEPAもLCMも、従来のLLMの弱点を補完・統合する方向性で一致しており、今後数年のAI研究における主要トレンドになると予想されます。

4. 人間の脳とLLMの類似性：神経科学から見た視点

LLMの振る舞いを論じる際、しばしば引き合いに出されるのが**「人間の脳は予測モデルである」**という現代神経科学の理論との類似性です。脳科学における**予測処理理論（Predictive Processing）**によれば、人間の脳は**常に外界からの感覚入力を予測し、実際の入力との差（予測誤差）を最小化するように働いている**とされます [51]。例えば私たちが文章を読むとき、脳は文脈から次に来る言葉を無意識に推測しています。予想外の単語が現れると脳波にN400と呼ばれる特徴的反応が生じることが知られていますが、これは脳が予測と違う入力に対して「誤差」を検出したサインと解釈できます [51]。この理論に基づけば、脳は一種の**「自己完結型シミュレータ」**であり、目や耳からの信号に先立って内的に次の状態をシミュレーションしているわけです。

この**脳の予測マシン仮説**と、LLMが行っている**次単語予測**との間には、表面的にも本質的にも類似点があります。実際、MITの研究では**言語タスクで高性能なAIモデルほど人間の脳活動パターンによく合致する**ことが示されました [52]。GPT系モデルのように次単語予測精度が高いモデルは、人間が文章を読んでいるときの脳の言語野（ブローカ野やウェルニッケ野）の活動パターンをかなり正確に予測できるのです。この発見は「知能という機能を実現する上で、生物の脳と人工のモデルが**類似の情報処理戦略**に収束しうる」可能性を示唆しています [53]。言い換えれば、**脳もLLMも「次に来るものを予測する」という共通原理で動いている**のかもしれません。

ごく最近の研究では、この類似性がさらに具体的に示されています。ヘブライ大学やGoogle Researchのチームによる2025年の研究では、**人間が音声言語を理解する際の脳内プロセスが、GPT-2やLlama2など大規模言語モデルの内部構造と驚くほど似ている**ことが報告されました [10]。被験者に30分間のポッドキャストを聞かせ、その間の脳活動（皮質脳波）を計測したところ、脳が話し言葉を処理していく時間的・階層的なパターンが、AIがテキストをトークン分解し文脈を積み重ねていく過程と対応していたのです [10]。具体的には、人間のブローカ野の活動が、モデル内部の**最初の層で単語をエンコードする処理**と同期し、次第に高次の言語理解に関わる脳活動パターンが**モデルの深い層での文脈統合処理**とシンクロするように現れたといいます [10] [54]。この結果は、「脳とAIは全く別物」という直感を揺るがし、**少なくとも言語処理という限定的な領域では、両者は驚くほど共通したアーキテクチャ的特徴を持つ**可能性を示しています。

もっとも、脳とLLMの間には**決定的な相違**も多数存在します。まず一つは**学習と適応の仕組み**です。LLMは一度大量データで事前学習された後は、そのパラメータ（重み）は固定され、追加の知識は微調整（ファインチューニング）や外部ツールの利用で補うしかありません。しかし**人間の脳は常に学習し続けます**。新しい経験はシナプス結合の重みを少しずつ変化させ、記憶痕跡として蓄積されていきます。言わば脳は**「常時オンライン学習中」**のモデルです。また脳には**明確な長期記憶・短期記憶の区別**があり、海馬を介したエピソード記憶の固定化など、時間スケールの異なるメモリシステムが統合されています。一方LLMの「記憶」は事前学習コーパスに由来する重み（長期記憶に相当）と、数千トークンのコンテキストウィンドウ（短期記憶に相当）に限られ、情報の更新・保持の柔軟性で脳に劣ります [17] [55]。

次に**知覚と身体性**の違いがあります。人間は視覚・聴覚・触覚・平衡感覚など**多感覚を統合**して世界モデルを形成しています [56]。「リンゴ」という概念を理解するにも、見た目、匂い、重さ、味、音（かじったときの音）といった**マルチモーダルな情報**が結びついています [56] [57]。さらに重要なのは、人間は**身体を持ち能動的に環境と相互作用する**中で因果や物理法則の直観を発達させる点です [58]。乳児がコップを掴んでは落とし転がし…という遊びを通じて「物は落ちる」「転がる」「ぶつかれば音がする」と学ぶように、**身体を介した経験（エンボディメント）**が人間のコモンセンス獲得の基盤にあります [58]。対照的に現在のLLM（および多くのAI）は**身体性を持ちません**。大量のテキストや画像から学習はしていますが、自ら行動して環境からフィードバックを得ることはないため、**経験に裏打ちされた直観**が欠けています [16] [39]。この違いは、「LLMがしばしば常識外れな誤答をするのは、物理的現実で試行錯誤した経験がないからだ」とする説明にもつながります [59]。近年、一部の研究者はロボット工学とLLMを結びつけ、物理環境でのインタラクションから学習させる試みを始めています [60]。将来的にはロボットが身体を持って世界を体験し、その知見を言語モデルにフィードバックするような学習が行われれば、このギャップも徐々に埋まっていくかもしれません [60]。

さらに**能動性（エージェンシー）の有無**も大きな違いです。人間の脳は予測するだけでなく、予測を踏まえて**行動を起こし**、その結果得られた感覚入力によって再び内部モデルを更新する、という**閉じた認知ループ**を回しています [61]。いわゆる**Active Inference（能動的推論）**の枠組みでは、生物は行動と知覚の両方で予測誤差を減らすように動くと説明されます [61]。一方、LLMは「言語的な出力」を返すだけで環境に直接作用しないため、このループが**閉じていない**との指摘があります [62]。確かに現状のLLMはユーザからプロンプトを与えられても自分から動くことはなく、外界からの追加入力も自発的には得られません。しかし興味深いことに、最近の論考では「LLMの出力テキストも広い意味では環境（人間社会）に影響を与えており、完全に受動的とも言い切れない。違いは**質的というより量的**ではないか」とする主張もあります [63] [62]。つまり、LLMに**継続的な対話や行動のフィードバック**を与える仕組みを組み込めば、より人間のActive Inference的な

性質に近づく可能性があります [63] [64] 。実際、最近登場した自己改善型エージェント（AutoGPT等）は、LLMにインターネット検索やコード実行といった行動手段を与え、結果をまたLLMにフィードバックするループを構成し始めています。これを発展させれば、LLMが**環境と相互作用しながら内部モデルを更新する**ような将来像も描けるでしょう。

最後に、**計算基盤の違い**にも触れておきます。人間の脳はニューロンのスパイク（発火）による**非同期並列計算**を行い、エネルギー効率は非常に高いです。一方、LLMは現代のデジタル計算機上で**同期的に行列演算**を行っており、演算効率や仕組みは脳とは大きく異なります。また脳には報酬系・感情・動機付けといった**認知以外の要素**も統合されており、これは現在のLLMには存在しません。例えば人間は報酬予測のためにドーパミン系が働き意思決定をバイアスしますが、LLMに「欲望」や「感情」はありません。同様に**意識や主観的体験の問題**も、人間の脳では未解明な重要テーマですが、LLMにはそもそもクオリアがあるのかといった問いも含め哲学的議論が盛んです [65] 。もっとも、これらは本質的な相違というより、現在のLLMが備えていない機能の列挙に過ぎず、今後AIに感情や動機を模倣させることが不可能と決まったわけではありません。

まとめると、人間の脳とLLMには**共通点**も**相違点**もあります。同じ「言語」という課題において、**予測モデル**としての共通原理が見られる一方で、**身体性・能動性・学習能の違い**が依然大きな隔たりとなっています。ただ近年の研究は、その隔たりを徐々に数値化・可視化し、どの方面から埋めていけばよいかを示し始めました。脳とAIの比較研究によって、AI側のモデル改良だけでなく、人間の知能の仕組み解明にも新たな示唆が得られるという**双方向の発展**が期待されています [53] [66] 。

5. 現行LLMの限界と今後3〜5年の技術的展望

現在のGPT-4やClaudeなど最先端のLLMが示す能力は驚異的ですが、その一方で**構造的な限界**もはっきりしてきました。前述したように、**①物理世界の理解不足、②永続的メモリの不在、③安定した推論の欠如、④複雑な計画の苦手さ**が主な弱点です [67] 。これらの制約を踏まえ、今後の研究開発は**LLMを超える新しいAIアーキテクチャや補完技術**に焦点が当たると見られます。MetaのLeCun氏は「現在のLLMパラダイムの賞味期限はあと3〜5年ほどで、誰もこれを中核には使わなくなるだろう。代わりに**新たなAIアーキテクチャのパラダイム**が台頭し、現行システムの限界を克服するだろう」と予測しています [68] [69] 。では具体的に、どのような進化が見込まれるでしょうか。

1. 世界知識と物理的常識の獲得：まず物理世界の理解については、前節で述べた**JEPAやマルチモーダル学習**の発展が鍵になるでしょう。テキストだけでなく**映像・音声・ロボットのセンサーデータ**など多様なモダリティから学習することで、AIにより**人間に近い常識的直観**を身につけさせる研究が進むと考えられます [39] [70] 。例えばGoogleやOpenAIは次世代モデルとして、画像・動画・音声を統合した**マルチモーダルLLM**（例：GoogleのGemini）はもちろん、ロボティクスと結合したAIエージェントの開発にも乗り出しています。今後3〜5年は、**「AIの身体性」**が一つのキーワードとなり、シミュレーション環境や実世界での試行から世界モデルを鍛える方向が加速するでしょう。

2.　長期記憶と継続学習：次に持続的メモリの問題については、**エージェントに経験を蓄積させるメモリフレームワーク**の研究が進展しています。例えば2025年に発表された**ReasoningBank**というフレームワークでは、LLMエージェントがタスクに挑戦する過程での**成功・失敗事例から一般的な戦略を抽出し記憶**する仕組みが提案されました [71] [72] 。従来、LLMエージェントは各タスクを**独立に**解いており、過去の失敗から学ぶことなく同じ間違いを繰り返す傾向がありました [55] 。ReasoningBankはこれを解決すべく、各試行の**教訓（成功パターン・失敗原因）を構造化したメモリ項目**として保存し、次回類似タスク時には参照して失敗を避けるようにしています [71] [72] 。例えば「商品の検索タスクであいまいなクエリを使うと候補が多すぎて失敗した」という経験から、「検索クエリを具体化しカテゴリ絞り込みする」という戦略を学習し、次の機会にはそれを活かす、といった具合です [73] [74] 。研究によれば、このような**自己記憶型エージェント**は従来のメモリなしエージェントよりもタスク成功率が向上し、徐々に**適応的で信頼性の高い**動作を示すようになったといいます [75] [76] 。今後は、このような**長期メモリを備えたLLM**が一般化し、ユーザと何度も対話する中でユーザの好みを学習したり、企業内でエージェントが業務知識を蓄積したりといった応用が現実味を帯び

るでしょう [77] [78] 。技術的には、知識の**エンコード・検索・更新**を効率よく行うメモリシステム（ベクトルデータベース+自己改善アルゴリズムなど）が重要になり、**LLMが逐次学習し続ける**仕組みが追求されると考えられます。

3. 推論力と計算能力の強化：LLMの論理推論力を高めるため、**モデルアーキテクチャの改良**や**自己監督による推論プロセスの明示化**が進むでしょう。既に**チェイン・オブ・ソート（CoT）**によってモデルに中間推論をさせる手法や、**自己教師ありでステップバイステップ思考を学習**させる取り組みが登場しています [79] 。OpenAIも「ツールの使用や中間思考をモデル自身に学習させる**プロセス監督**」を研究しており、将来的にはモデルが内部に**論理演算モジュール**や**計算モジュール**を持つような形に発展する可能性があります [80] 。また、現在はソフトウェア的にLLMを補佐する形で**ツール使用エージェント**（電卓や検索エンジンとの連携）が実現していますが、この連携をモデル内部に巻き取る、すなわち**「計算できるLLM」**や**「検索機能付きLLM」**といった拡張も考えられます。Googleの次世代モデルではコード実行能力の統合が示唆されていますし、学術界でも言語モデルに微分方程式ソルバー等を組み込む試みがあります。さらに、推論の安定性向上には**自己評価と反省**の仕組みも有効とされています。Anthropic社はClaudeに**自己検閲・自己反省**のプロンプトを用いて一貫性を高める研究を報告しており、モデル自身が**「答えの妥当性を再評価する二段構え」**のような推論手法も広まるでしょう。

4. 計画・行動制御の導入：複雑なタスクをこなすためには、LLMに**プランニング能力**を付与する必要があります。今後数年で、LLMを中核としつつも外部環境に対してループを閉じた**AIエージェント**が実用化に近づくと予想されます [81] [64] 。具体的には、LLMがサブゴールを設定し、逐次外界に働きかけ、結果を読み取って方針を修正する――この一連を自律的に行うシステムです。すでにAutoGPTやBabyAGIといったプロトタイプが登場し話題を呼びましたが、2025年現在の時点では信頼性や効率の面で実用には課題があります。しかし、例えば前述のメモリ機構と組み合わせて過去のタスク経験から方略を学習したり、あるいは**階層型のエージェント構造**（高レベル計画担当と低レベル実行担当を分けるなど）を採用することで、**ゴール指向の安定したタスク遂行**が可能なAIが登場するかもしれません [82] [50] 。LeCun氏が言うように「次のパラダイム」ではLLMは中核コンポーネントではなくなるかもしれませんが、それはすなわち**LLMが大きなシステムの一部（計画モジュール・推論モジュール・対話モジュール等の）となる**ことを意味します [69] 。今後3〜5年は、その統合アーキテクチャの試行錯誤が進み、特に**ロボティクス分野との結合**によって**「現実世界で動くGPT」**が出現する可能性もあります。実際、「これからの10年はロボティクスの10年になる」との予測もあり [83] [84] 、物理世界で活動する知的エージェントへのLLM応用が大きなテーマとなるでしょう。

5. 信頼性・安全性・倫理の確保：技術的進歩と並行して、LLMの**安全性と倫理**の問題も引き続き重要課題です。現行のLLMは有害発言の抑制やバイアス低減のためにRLHF（人間フィードバックによる強化学習）で調整されていますが、将来のより強力なモデルでは一層高度な**アライメント（価値観の整合）技術**が求められるでしょう [85] 。開発コミュニティでは、モデルに自己検閲させる「コンスティテューションAI」や、第三者機関がモデルを評価する仕組みなど、様々な対策が議論されています。技術的には**モデルの内部決定プロセスを説明可能にする**工夫（前述のLCMのようなアプローチ）も安全性向上に寄与します [43] [47] 。今後数年で、法律や規制の整備も進み、AIモデルの透明性やテスト手法が標準化されていく可能性があります。技術面でも**「フェイルセーフなモデル」**つまり間違えても暴走しない・説明責任を果たせるAIへのニーズが高まり、それが研究開発の方向性にも影響を与えるでしょう。

以上を踏まえ、**今後3〜5年の展望**としては、「LLM単体の精度競争」から「LLMを含む統合知能システムの構築」へと焦点が移っていくと考えられます [68] [67] 。LLMはこれまで、人間の言語という**知的活動の断片**を驚くべき精度で模倣するところまで来ました。次のステップでは、それを**世界全体のモデル**へ拡張し、**記憶し・考え・計画し・行動する**総合知に高める挑戦が始まっています [50] 。それは決して容易な道ではありませんが、着実に萌芽的な成果が現れつつあります。おそらく今後数年のうちに、「あのChatGPT登場以降停滞していたと思われたAIが、再び質的飛躍を遂げる瞬間」が訪れるでしょう。その原動力となるのが、ここで述べた**世界モデルの導入**であり、**人間の認知機能から学んだ新原理の統合**なのです [86] 。研究者たちは既にその方向に向かって動き始めており、LLMの次章がどのように描かれるのか、注視されます。

参考文献・出典：

- 【8】 SingularityNET Ambassadors, “Beyond LLMs: Memory, Reasoning and ‘True’ Generative Models,” Medium, Dec. 2025 87 11 .
- 【12】 ASIに仕事を奪われたい, 「MetaのAI天才がついに退職――...」, 2025 27 .
- 【13】 ASIに仕事を奪われたい, 同上 14 12 .
- 【14】 ジコログ, 「LLMは本当に行き止まりか？...技術論争」, 2025 16 22 .
- 【16】 Jacob Andreas, “Language Models, World Models, and Human Model-Building,” MIT Blog, 2024 9 .
- 【18】 Neel Nanda, “Othello-GPT Has A Linear Emergent World Representation,” 2023 4 5 .
- 【22】 AI Alignment Forum, “Knowledge Neurons in Pretrained Transformers,” 2021 2 .
- 【23】 Anthropic, “Tracing the thoughts of a large language model,” 2024 80 .
- 【26】 Anil Ananthaswamy, “This AI Model Can Intuit How the Physical World Works,” WIRED/ Quanta, Dec. 2025 29 32 .
- 【29】 Anidhya Bhatnagar, “Large Concept Models: Paradigm Shift in AI Reasoning,” InfoQ, 2025 46 44 .
- 【31】 BioErrorLog, 「LLMと脳理論: Active Inferenceの違いと類似点」, 2024 63 62 .
- 【32】 中小企業AI活用協会, 「人間の脳とAIの知能メカニズム比較」, 2025 53 51 .
- 【33】 innovaTopia, 「脳はAIのように言語を理解する――Nature誌...」, 2025 10 .
- 【35】 Paul Sawers, “Yann LeCun... new paradigm of AI architectures,” TechCrunch, Jan. 2025 67 50 .
- 【36】 Anil Ananthaswamy, “How One AI Model Creates a Physical Intuition...,” Quanta Magazine, Oct. 2025 30 .
- 【40】 Ben Dickson, “New memory framework builds AI agents...,” VentureBeat, Oct. 2025 55 71 .

---

1 11 18 19 20 25 87 Beyond LLMs: Memory, Reasoning and “True” Generative Models | by SingularityNET Ambassadors | Dec, 2025 | Medium
https://medium.com/@singularitynetambassadors/beyond-llms-memory-reasoning-and-true-generative-models-e0ca6da6d1f7

2 Knowledge Neurons in Pretrained Transformers — AI Alignment Forum
https://www.alignmentforum.org/posts/LdoKzGom7gPLqEZyQ/knowledge-neurons-in-pretrained-transformers

3 4 5 6 7 Actually, Othello-GPT Has A Linear Emergent World Representation — Neel Nanda
https://www.neelnanda.io/mechanistic-interpretability/othello

8 9 Language Models, World Models, and Human Model-Building
https://lingo.csail.mit.edu/blog/world_models/

10 54 66 脳はAIのように言語を理解する――Nature誌掲載、ブローカ野と大規模言語モデルの驚くべき一致
https://innovatopia.jp/neurotrechnology/neurotrechnology-news/74067/

12 13 14 27 MetaのAI天才がついに退職――ザッカーバーグも驚いた模様 | ASIに仕事を奪われたい
https://asi.tokyo/2025/11/13/meta%E3%81%AEai%E5%A4%A9%E6%89%8D%E3%81%8C%E3%81%A4%E3%81%84%E3%81%AB%E9%80%80%E8%81%B7%E2%80%95%E2%8

15 50 67 68 69 81 82 83 84 86 Meta's Yann LeCun predicts 'new paradigm of AI architectures' within 5 years and 'decade of robotics' | TechCrunch
https://techcrunch.com/2025/01/23/metas-yann-lecun-predicts-a-new-ai-architectures-paradigm-within-5-years-and-decade-of-robotics/

16 17 21 22 23 24 26 39 40 41 59 LLMは本当に行き止まりか？Meta研究者の発言が引き起こした技術論争 | ジコログ
https://self-development.info/
llm%E3%81%AF%E6%9C%AC%E5%BD%93%E3%81%AB%E8%A1%8C%E3%81%8D%E6%AD%A2%E3%81%BE%E3%82%8A%E3%81%8B%EF%BC%

28 29 31 32 33 34 35 This AI Model Can Intuit How the Physical World Works | WIRED
https://www.wired.com/story/how-one-ai-model-creates-a-physical-intuition-of-its-environment/

30 36 37 38 How One AI Model Creates a Physical Intuition of Its Environment | Quanta Magazine
https://www.quantamagazine.org/how-one-ai-model-creates-a-physical-intuition-of-its-environment-20251003/

42 43 44 45 46 47 48 Large Concept Models: a Paradigm Shift in AI Reasoning - InfoQ
https://www.infoq.com/articles/lcm-paradigm-shift-ai-reasoning/

49 51 52 53 56 57 58 60 70 人間の脳とAIの知能メカニズム比較｜予測処理理論から読み解く知能の未来
https://research.smeai.org/ai-brain-intelligence-comparison/

55 71 72 73 74 75 76 77 78 New memory framework builds AI agents that can handle the real world's unpredictability | VentureBeat
https://venturebeat.com/ai/new-memory-framework-builds-ai-agents-that-can-handle-the-real-worlds

61 62 63 64 LLMと脳理論: Active Inferenceの違いと類似点 - BioErrorLog Tech Blog
https://www.bioerrorlog.work/entry/active-inference-llm-predictive-minds

65 ChatGPTの思考回路と人間の脳の違いを徹底解説 - WEEL
https://weel.co.jp/media/innovator/ai-human-comparison/

79 State of Reasoning LLMs: The New Era of “Thinking” Machines
https://medium.com/@adnanmasood/state-of-reasoning-llms-the-new-era-of-thinking-machines-f241b1a3096d

80 Learning to reason with LLMs | OpenAI
https://openai.com/index/learning-to-reason-with-llms/

85 Turing award winner says llms are fundamentally flawed - Facebook
https://www.facebook.com/groups/aifire.co/posts/1819237282014906/