今、破壊的 AI について考えるべきこと

AI の進化、脅威への対応

Exscale Inc. Founder CTO
萩原正義

ChatGPT の出現以降、急速に AI 技術が進化している。AGI あるいは Super intelligence の出現も近いと言われている。我々は今、何に注目し、考えていくべきか。IT 技術の実践面だけでなく、認知科学、哲学、社会学などの幅広い観点での議論と、AI の進化、脅威への対応を提言する。

キーワード： AI、LLM、AGI、Super intelligence

1　はじめに

現在の AI の進化はある程度予想されていたが、ChatGPT などの Large Language Model(LLM)のサービス利用が開始されてから破壊的な進化が見られている。今後、クラウドコンピューティングの計算能力の過半数が AI の学習、データ生成に使われると予想されている。この LLM の進化では、単語の意味を embedding（埋め込み）[1]と呼ばれる多次元のベクトル空間で表現する技術と、これまで教師あり学習で見られた、学習に使われるデータセットへ人手によるラベル付けを不要とする自

己教師あり学習[2]の技術基盤の寄与が大きい。

embedding とは、これまでの画像などのオブジェクトの識別に複数の属性（カラム）を組み合わせたパターンを使う方法を拡張し、自然言語の単語の意味の識別に必要十分な数の属性の次元を用意して、その次元で意味を識別するベクトル表現である。その表現を用いてモデルを学習する。有名なものに Word2Vec[3]がある。自己教師あり学習とは、学習対象の文の一部の単語をマスクして隠して、その前後の文脈から隠した単語を予測し、その正解を学習していく方法である。元のデータセットをお手本に使うことで、人によるラベル付けが不要になり、学習に必要なデータセットの準備や前処理に人手が不要となって自動化できる。受験勉強で、参考書の一部をマスクして暗記を繰り返す学習法と原理的には同じである。

学習の自動化と並んで重要なのは倫理の学習である。学習の結果、倫理的に問題がない回答を確実に得るには、従来は強化学習によって人手で回答の倫理性を判定していた。その後、人手によらず回答の倫理性の判定を自動的に学習させる手法、DPO（Direct Preference Optimization）[4]が発明され、効率化の問題が解決しつつある。DPO はあらかじめ倫理性の問題となる回答として、一方は倫理的に正しい、他方を倫理的に問題があるデータセットを用意して学習させる。Preference とは、好ましさに関するラベル付けで、倫理やその他の領域での善悪や美醜などの判断指標を与える。実際の判定は、確率的な推論になる。この判断はラベル付けを行う人間の主観がバイアスに含まれるので、どのように客観化するかの問題は解決できない。

現在の LLM は主に自然言語に対応がとどまるが、画像、音声、動画に拡張してそれらの認識（識別や分類など）に応用するマルチモーダル対応と、１つの LLM をトピック分類、機械翻訳、意思決定など複数のタスクに応用するマルチタスク対応が進められている。このような急激な進化の後には、AGI あるいは Super intelligence と呼ばれる人を超越する能力を持つようになるだろう、と言われているが果たして本当だろうか？

2　現在の AI の課題

2－1　原理的な問題

すでに AI は人の能力の一部を超越している。たとえば、オブジェクト検出や画像識別では人の認識能力を上回る例は多いし、Bonanza[5]、AlphaGo[6]などは対戦型ゲームで人を圧倒している。ただし、こうした能力は人の多彩な認知機能の能力の一部である。人の認知能力には記憶力、言語能力、判断力、計算力、遂行力の 5

種類が存在する。LLM はたくみに文を構成し文書を生成するが、簡単な推論に誤りが生じる場合や、複雑なタスクの実行はまだできていない。これらの問題は、既存のデータベースや semantic search[7]、knowledge graph[8]、強化学習などとの組み合わせによって改善の途中にある。

人の認知機能に関して、認知科学では 4E の特性があると言われている[9]。4E とは、Embodied, Embedded, Extender, Enactive の 4 つの頭文字を示している。たとえば、人は電卓で計算機能の一部を代行させたり、メモを取ることで記憶を外部に記録として残す。これらの行動は、人の持つ認知機能である計算力や記憶力を外部ツールと連携して拡張している例で、Extender で示された認知機能である。LLM を使うアプリケーションのアーキテクチャでも、図 1 に見られるように、外部データベース（Vectorstore）に文書の文のまとまりの embedding 表現を記憶させることで、LLM の記憶機能に Extender を応用する方法が一般化している[10, 11]。このように認知機能を拡張するアプローチは LLM や AI の発展で続くものと予想する。Embodied に関しては、AI の永年の問題であった記号接地問題[12]に関係する。記号接地問題は、現在の LLM が言葉の意味を理解していないという批判の元になっている。LLM は単なる「次の単語予想機」で意味を考えて文章を生成していないという指摘である。たとえば、りんごの意味は、見た目だけでなく、味、香り、食べたときの歯ざわりなどの感覚が一体になって初めて意味が認識される。言葉の意味の理解ではこの感覚の裏付け、身体的な感覚を持つこと、記号接地の概念が必要になる。LLM は自然言語の言葉しか対象としていないので、身体も感覚器官も持たず、記号接地問題は解決できない。すなわち、言葉の意味の理解は不可能である。

（注）ただし、この説明には人はラベル（言葉）を通じて意味を認識する前提がある。教師あり学習でデータセットに正解ラベルを付けるのも、ラベルの言葉を通じてデータセットのオブジェクトを認識する人の認知機能の前提がある。

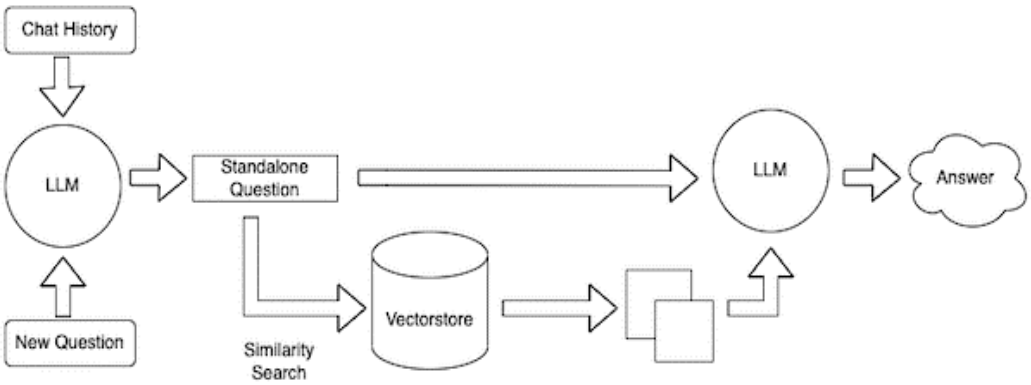
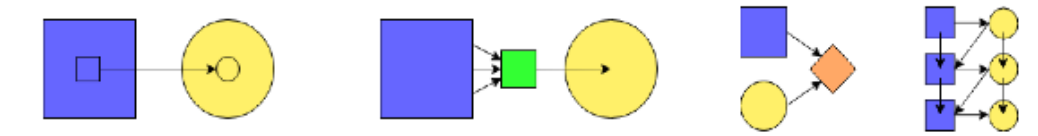


図 1　外部ツールを使った LLM アプリケーション

チャットでユーザからの新規の質問がなされると、Vectorstore から embedding を使って類似検索を行って回答を得る。その回答をヒントに LLM に質問とともにプロンプトを入力し、質問への回答を生成する。

（出典： Tutorial: ChatGPT Over Your Data）

前述のように、LLM は自然言語以外にも、画像や動画の認識などマルチモーダル化の応用が進んでいる。LLM はすでに画像を説明する文章の生成や、動画を認識してシーンを文章で説明可能となっている[13]。この能力を進めていけば、視覚、聴覚、ロボットアームのような触覚をベクトル表現で識別することで、「総合的」に状況を把握することが可能となるだろう。つまり、身体的な感覚に基づいた認知機能を持つことになる。言葉、画像などのオブジェクト間の意味的な関係性、依存関係を認識することで、関連する意味、概念のまとまりを押さえることが可能となり、いわゆるフレーム問題[14]も解決に向かうだろう。同時に説明可能性を与えることにもなる。図 2 は 2 モーダル間、例えば、言語と画像を組み合わせた認識の例を示している[15]。



(a) Cross-modal transfer. Information from modality *A* is aligned to comparable information in *B*.

(b) Cross-modal interpretation. Relevant information in modality *A* is summarized and interpreted in modality *B*.

(c) Joint multimodal processing. Left: Modality *A* and *B* both contribute to a joint prediction. Right: Interactive exchange of information between modalities.

図 2　マルチモーダルによる記号接地問題への対応

マルチモーダルタスクでの情報の流れ。青と黄色はモーダル A とモーダル B を示す。

(a)モーダル A からの情報は対応するモーダル B の情報に対応する。モーダル間の転送。

(b)モーダル A の適切な情報がモーダル B で要約され解釈される。モーダル間の解釈。

(c)左はモーダル A、B を合わせた予測の実行、右はモーダル間で相互作用的な情報交換。マルチモーダルの結合処理。

（出典： Multimodal Grounding for Language Processing）

オブジェクト間の意味的な関係性、依存関係の原理的な認識に関して、ニューラルネットワークは複合化の推論ができない古典的な問題が 30 年来存在している。たとえば、未知の単語の規則性を推論し、新たな未知の単語にその規則性を当てはめて定義すること、あるいは、未知の単語の組み合わせの規則性を推論し、新たな未知の単語の組み合わせにその規則性を当てはめて定義すること、の推論ができていない。しかし、少ない例を学習させ、単語の複合化の推論が言語モデルで可能となるメタ学習アプローチが提案されている[16]。これは人の言語や思考能力の基盤となりうるだろう。

それでは、AI に残された課題は何か。

たとえば、認知機能以外の意識はどうだろうか。自己認識もこの一部となるだろう。あるいは、正義、節度、勇気、倫理観など、いわゆる存在のレルムも考えていく必要がある。ただ、これらは、哲学、心理学、倫理学、社会学、語用論などにかかわる広範な問題で、AI の専門家だけで対応できる問題ではない。

解決のヒントになる技術と概念としては、認知アーキテクチャ[17]、ニューラルシンボリック AI[18]などの技術要素や、オントロジー、二元論(Dualism)[19]、ダニエルカーネマンの二重過程理論(Dual process theory)[20]、ロバートキーガンの成人発達理論[21]などの思想や理論がある。

2－2　AI アプリケーション開発上の問題

AGI あるいは Super intelligence の実現の高度な問題ではなく、AI アプリケーション開発上のより実践的な問題を考えてみよう。筆者が start up で製品開発する上で現状の LLM がうまく扱えない問題ととらえているものは以下である。

- 命題と非命題の区別
- 真実と虚偽（フェイク）の識別
- 矛盾点の発見
- 事実と解釈の区別
- 主観と客観の区別
- 相関と因果の区別
- 本音と建前の区別
- 意図(インテント）の識別
- 抽象と具体の対比、行き来
- 意味の粒度（スコープ）の対比
- 仮説（アブダクション）や公理の設定

これらの課題を解決する論文は現状ではどれもあまり原理的な解決法となりえて

いない。このあたりが AI の原理的な問題とは別に、実践的な開発上の差別化になっている。

3　今後の AI の進め方[22, 23]

AI 関連技術の進化により開発コストは低下し、AI 利用者も増大するので、今のうちに AI の潜在的な課題に対する対策が必要である。

AI の課題を解決し、その進化を加速するために以下を考える。

（1）開発の自動化

現在の LLM の開発では、RLHF(Reinforcement Learning from Human Feedback：人間のフィードバックからの強化学習)を学習に用いている。この方法は、人間が学習結果を確認することで学習内容の適切さを維持する、いわゆる試行錯誤である。しかし、人手で実行することで、スケールしないこと、人の育成が必要であること、監視能力に限界があることなどの課題を持っている。また、人による評価が困難な作業も存在する。文書の要約が適切か、事実が正しいか、学習結果の振る舞いが維持されるか（ロバストネス)、敵対的テストに学習パイプライン全体が耐えうるかなどである。これを自動化するための大規模な学習方法、検証方法、負荷テストの構築が求められる。

(2) 評価法の確立[24]

AI の社会的影響は AI システムの真実性、公平性、誤った利用の可能性などの品質に依存する。したがって、AI システムは「評価駆動型」の開発法を採用すべきである。ここで現在の評価法とその現状課題を議論しておく。

- 複数選択式評価: 数学、歴史、法律など 57 タスクからなる精度評価 Massive Multitask Language Understanding (MMLU) や 9 の社会的バイアスを評価する Bias Benchmark for QA (BBQ)
- サードパーティ評価フレームワーク: 科学から社会的思考までのボトムアップ評価 BIG-bench や専門家によるトップダウン評価 Stanford's Holistic Evaluation of Language Models (HELM)
- クラウドソーシング A/B テスト: モデルの有効性や有害性によるランキング。人による評価は主観的で人の特性に依存する
- ドメイン専門家評価: 国家セキュリティに関する評価

- 生成 AI による評価法の開発: 人による精度確認を伴い、社会的バイアスは排除できない
- 外部監査機関: 監査の健全性を維持するため、監査方法を明かさないことが必要だが、それにより開発側は評価法の改善が困難となる課題を持つ

（3）ルール決定の民主的プロセスの構築

LLM に対する質問への回答では、たとえば、質問者個人の好みや価値観に従う程度、共感的支援の機会、公表データへの見解における中立性、医療・金融・法律関係の回答のための条件、言語と画像を組み合わせた認知での人物の識別の程度、人物画像の生成におけるあいまいな要求に対してどの属性を優先して生成するか、LGBTQ や女性権利など人権、文化の地域による違いの考慮した生成の制限や拒否のためのコンテンツ分類基準、などを決める必要がある。

また、これらの条件、品質基準を決める評価メトリックスを掲示すべきである。

多様な背景や AI の理解度を持つ個人をどう参加させ、少数派意見を取り組む機会を提供するか、幅広い見解から有益な結果を得るためのフィルター調整が公平かどうかも考えていく。

手間のかかる個別参加型ではなく、スケールする仮想的プロセスの実行可能性、プロセス自体の透明性、信ぴょう性が容易に理解できることが条件となる。参加者の選出方法では、少数派意見者として子供向けの教育者、心理学者、地域的ポリシーに関しては地域住民の参加を考える。参加者の相互理解、意見の違いに関する深い理解、意見調整のプロセスを含めて、代表者の選出、多数決やその他の投票方法、非作為サンプル、仕分けの方法などの採決法を検討する。

（4）ガバナンススコープ

LLM におけるある能力以上の取り組みには IAEA のような査察機関を設置し、安全性への適合テスト、展開やセキュリティレベルへの制約を加え、計算量やエネルギー利用を追跡する。ただし、制約を加え過ぎないように、国家の慣習や文化に依存した制約は各国にゆだね、その以外は自由な開発を認める。生み出す価値がリスクに見合うことが重要である。

制約レベルは実験的に確立する。その制約のもとでは、個人が技術を制御できることが重要となる。

4 提言：AI の方向性のブループリント

LLM は直近では GPTs や virtual assistant といったエージェントへの進化が見られる。エージェントは、これまでのプロンプトによるユーザからの指示を待たずに、自らがユーザの意図を推論して自律的に行動するソフトウェアである。ロボットのようなハードウェア形状を伴う場合もある。内部に行動規範を作成し、また、ある程度、意図や行動に対する自己修正をする。これにより、ユーザが何らかの目的で作業を実行したい場合、その目的から意図する行動をエージェントはユーザに代わり実行する。これは人間が AI に面倒な作業の実行や手順を踏んだ複数の操作の自動化を期待する必然的進化の方向性である。さらに次の段階では、複数のエージェントが協調し、人間のチームの一員として動作することになるだろう。現在は、まだマルチエージェントによる実行可能な作業は限定的であるが。

我々はこうした環境において必要となるアーキテクチャ、要素技術、解決すべき課題を議論している。以下ではその概要を説明する。

4 − 1 組織論

エージェントと人が協調するチーム（超マシーン[25]と呼ぶ）は、特定の目的に従って構成される。チームの目的は当初から決められる場合と、漠然とした目標があり活動をしながら目的を定義し修正していく場合とがある。混沌とした現在では、後者の場合がより実践的であろう。チームは目的の達成に必要なタスクを抽出し、構成メンバーの能力に応じてそのタスクを分担する。これまでのプロジェクト管理ではメンバーの役割に応じてタスクを分担していたが、ここでは、限られた制約（納期、コスト、設備、情報など）の下で、メンバーの能力という有限リソースをタスクに分配する計算可能な実装法をとる。これをセル組織[26]と呼ぶ。セルとは自律的に動くチーム単位である。エージェントは LLM が実行するので、汎用性の高い能力を有するが、特定ドメイン向けファインチューニングやドメイン依存の外部知識を参照することで特定の能力を持つことになる。

4 − 2 行動規範

エージェントの意図しない、不正な行動を抑制するために、従来の LLM 向けのセキュリティ対策が施される。エージェントの倫理的振る舞いの学習法では、具体的な個別の行動の善悪、真偽の判定を教えるだけでは、抜け漏れが生じて対処療法的でかつ非効率である。人的負担も増大してしまう。したがって、人権保護などの原則を基本とし、それを補完する事例とともに学習させる体系化が必要となる。つまり、ルールベースから原則ベースへの考え方の転換である。

企業や組織では組織ぐるみの大きな不正が発生する可能性もある。したがって、組

織構造を相互に監視し抑制する、三権分立の構成が必要となる。これはビザンチン問題の解決プロトコルとして一般化可能である。

エージェントや人の自律性は、それぞれの自由が最大限尊重されるべきであるが、社会全体や組織の構造の制約条件が必ず伴う。エージェントの自律性は、自動運転に見られる AI の自律性の分類[27]が参考になる。

0 自動制御なし、1 単純な自動化（定常運転、例外事象を想定しない)、2 部分的自動化（基本的機能の自動化、人による監視と人への制御の切り替え常時可能性)、3 条件付き自動化（環境の現象監視と大部分の機能の自動化、人への制御の切り替え)、4 高度な自動化（特定環境下で全機能の自動化、人への制御の切り替えはオプション)、5 全自動化（全環境下での全機能の自動化、人の制御不要)。

LLM のエージェント性[28]とは、人による部分的な管理下において、複雑な目標を、複雑な環境で、適応的に達成する度合いと定義される。エージェント性は、達成可能な目標の範囲や限界を示す目標の複雑さ、目標を達成するための環境の複雑さ、新しい状況や想定外の状況への対応能力を示す適応性、人の介入や監視の必要性の度合いを示す実行独立性、の 4 要素で構成される。この自律性に対して、社会や組織の制約が加えられる。

トマス・アクイナスの人格と倫理学[29]では、人格は自由な存在であるとし、自由な意志で選択した倫理的行為を行い、その行為に責任を負う。自己所有、自己支配的な人格を実現するのが意志で、その意志による行為を発動させるのが理性である。この自己決定を自律の意義とする。このトマス倫理学に基づいた補完性原理 the principle of subsidiarity では、人の自律性に対する社会の制約のあるべき姿を、個人の目標達成を尊重し、その支援を社会が実行するが、社会は個人が自力で行えることまで干渉してはいけないとしている。自律性は、意志を持ち倫理的な行為に責任を持つ理性を前提とする。

4 － 3 世界モデル

エージェントは、外部環境の現象を観測し、認識し、解釈し、判断して、アクションを決定する。言語（で示される概念）を通じて世界を見る。これは、二重過程理論でのシステム 2 の動作である。一方、緊急性の高い行動が必要の場合は、観測、認識、アクションの段階を踏むシステム 1 の動作となるべきである。認識、アクションの間には直観に近い判断が含まれるかもしれない。

（式 1） 現象－観測－認識－（解釈－判断）－アクション

もし、エージェントが人間とチームを組み、言葉によるコミュニケーションを取る場合、人間と同じくエージェントにこのシステム 2 / 1 の動作を仮定することは現実的であろう。ここで、このコミュニケーションでは、AI の課題であるフレーム知識や記号接地問題がある程度解決できたとしても、能力的には十分と言えない。エージェントが現象を認識し、解釈する上で人間が持つメンタルモデルに相当する世界モデルが必要となるからである。世界モデルとは、現在の状態で欠けている情報の推定や状況の不確実性や状況の変化に対して俯瞰的で予見的な説明を与えたり、過去の現象から未来の現象の発生を予測する因果関係の把握をしたり、一部から全体像を推論する能力を生み出す世界像である。それによって将来のシナリオを想定し事前の戦略の準備と調整を行う。人と類似の認知プロセスや意思決定の機能を与える。

世界モデルの例として、RNN、SSM、RSSM の 3 つを挙げる[30]。ある時点の潜在変数空間表現（意味の場）の状態からアクションの実行を通じて別の状態に移る過程をモデル化し将来の状態の予測を与えるモデルを考える。確実性の高い要素と不確実性のある要素を分離して将来を予測することで、情報の継続性を確保した学習能力（RNN の強み）と不確実性に対応する高い予測能力（SSM の強み）を両立するハイブリッドモデル RSSM が与えられる。

人は現象を観測し解釈を通じて主観的にパターンを推論して学習する。それらの学習結果が総体として主観的な世界を形成する。そこでは、物理的な法則や慣習、道徳、倫理といった学習済みの背景知識の影響を受けて因果関係を形成する。たとえば、AI で車が走行して砂煙を上げるシーンや木葉が風でなびくシーンを模倣することで、あたかもそれらの動きを理解しているかの表現が可能である。しかし、あくまでも模倣であるため、ガラスに石が当たってガラスが割れるシーンの予想はできない。なぜなら、石が当たることで力が加わってガラスが割れる物理法則を学習しているわけではないから。また、電車の乗り降りで乗る人が降りる人を待つシーンの行動も予想はできない。道徳的に降りる人に道を譲ることが習慣となっている背景が理解できないからである。こうした物理法則の理論や道徳、倫理の見えない理由によって引き起こされる因果関係を模倣から導く段階にはまだ AI は至っていない。ＡＩにも因果発見や因果推論の分野が存在するが、それを LLM の学習に結び付けて内部で表現学習し、世界像を形成することはできていない。AI の因果推論では、因果関係の存在の可能性を示すだけで、その因果関係が生じている理由は示さない。その理由を AI が認識可能となるには、因果関係を生じさせる物理法則や倫理、習慣を含めた全体的な関係の発見が必要になるからである。

（式２） 現象（出来事やアクション間のパターン）－模倣－（理論と倫理）－因果関係

AI のうち特に深層学習では概念（言葉、画像）や現象の入力に対して、潜在変数空間表現（意味の場）が割り当てられる。この表現の割り当てを決める深層学習を構成するネットワーク構造の重みづけが学習のたびに修正される。エージェントはそれぞれが自律的に別々の表現を学習するので、独自の潜在変数空間表現を持っていると見なせる。そのうえで、出力の言葉の意味定義の範囲内で互いにコミュニケーションを取れることでマルチエージェントの動作を可能としている。

エージェントが人と社会で共存していくためには、システム 2 / 1 の認知機能に加えて世界モデルが人の世界像と類似しなければならない。ここで参考になるのが、Penrose 氏の 3 世界モデル[31]である。このモデルは、現実世界の現象が存在する前提で、その一部が認識され、さらにその認識の一部が（物理や数学で）理論化される。さらに、その理論の一部が現実世界にマップされて、1 周循環する。Wolfram 氏の Principle of Computational Equivalence （計算等価性の原則）[32]もこれに類似している。あらゆる物理現象はオートマトン理論により説明可能としている。科学領域の様々な現象は，実はその根底において同一のアルゴリズムに支配されており，これを色々なやり方で反復計算することによって，各領域の複雑な現象が生成されるという。両者のモデルでは、現実世界の現象は仮に認識、理論化できなくても存在は認めている。認識されても理論化できない現象や、理論化されても現実世界の現象として生じないものも存在を認めている。世界を認識できるものだけに限定する認識論を超えた超越的な立場であり、また、理論化できない現象の存在を認める実在論の立場でもある。

なぜ、エージェントの世界モデルもこの立場になるかと言えば、エージェントの動作が引き起こす相転移や、社会との相互作用による活動が複雑系となっていて、理論的な十分な解明が不可能である一方で、有益な活動を機能させるからである。複雑系の持つ現象を完全に説明可能でなくてもその存在は認める立場は取りたい。また、エージェントの組織的協調活動、エージェントの行動規範と社会の倫理規範、法律、さらに、経済活動、外交、教育、科学の進化などを全体システムと見た場合、それら全体を計算可能ととらえたい。もちろん、できるだけ動作を説明可能とし、自己破滅しないよう安定的な平衡状態を取りながら進化を可能としたい。計算可能な表現を取りつつも、完全な理論化はせず、計算的解決が困難な問題ととらえたい。

この立場に立つに至った理由は、LLM の振る舞いが、自然言語だけを学習した場合に比べて、プログラミング言語の学習を加えた場合に、性能の改善が実証されたからである[33] [34]。

たとえば、問題を Chain of Thought[35]という手法で小問題の手順に分解し、それぞれの小問題をプログラムコード化して回答を出す例[36]が存在する。プログラム

言語によって世の中の一般的な問題を実行可能としている初歩的な例である。

また、プログラム言語を自然言語として見なして実行する例[37]がある。プログラム言語自体は当然実行可能であるが、これを自然言語の一種とみなす。プログラム言語を学習した LLM が高度な性能を獲得した結果、逆に自然言語をプログラム言語のように実行可能としてしまうことで、世の中の問題を命題や制約付与だけで、ルールやプログラムで記述しなくても実行可能とする。これで自然言語の実行可能性が一歩近づく。

ここで、計算可能性では、Wolfram 氏のモデルにようにチューニングマシンやオートマトンを使った計算モデル化を想定している。ただし、停止問題は計算モデル内での解決に限定せず、人による適切な解決を考える。そして、複雑系であるシステム全体を知識集積の場とし、エージェントと人が並列実行する計算可能なモデルで表現される意味の場と考える。システムを計算可能な表現とすると、エージェントの性能にも改善が図れるのは前述した通りである。

参考文献

[1] Embedding 単語の埋め込み - Wikipedia

[2] 自己教師あり学習 自己教師あり学習 - Wikipedia

[3] Word2Vecc Word2vec - Wikipedia

[4] Direct Preference Optimization: Your Language Model is Secretly a Reward Model https://arxiv.org/abs/2305.18290

[5] Bonanza https://ja.wikipedia.org/wiki/Bonanza

[6] AlphaGo https://ja.wikipedia.org/wiki/AlphaGo

[7] Semantic search https://en.wikipedia.org/wiki/Semantic_search

[8] Knowledge graph https://en.wikipedia.org/wiki/Knowledge_graph

[9] The Oxford Handbook of 4E COGNITION https://academic.oup.com/edited-volume/28083

[10] Tutorial: ChatGPT Over Your Data https://blog.langchain.dev/tutorial-chatgpt-over-your-data/

[11] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

https://arxiv.org/abs/2005.11401

[12] The symbol grounding problem
https://eprints.soton.ac.uk/250382/1/symgro.pdf

[13] Multimodal Chain-of-Thought Reasoning in Language Models
https://arxiv.org/abs/2302.00923

[14] Frame problem フレーム問題 - Wikipedia

[15] Multimodal Grounding for Language Processing
https://aclanthology.org/C18-1197/

[16] Human-like systematic generalization through a meta-learning neural network

https://www.nature.com/articles/s41586-023-06668-3

[17] Cognitive architectures Research issues and challenges
https://www.sciencedirect.com/science/article/abs/pii/S1389041708000557

[18] Neuro-symbolic AI https://en.wikipedia.org/wiki/Neuro-symbolic_AI

[19] 2 元論 二元論 - Wikipedia

[20] Kahneman, Daniel Thinking, Fast and Slow
https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow

[21] Robert Kegan The Evolving Self

[22] An Overview of Catastrophic AI Risks https://arxiv.org/abs/2306.12001

[23] OpenAI Governance of superintelligence
https://openai.com/blog/governance-of-superintelligence

[24] Challenges in evaluating AI systems

https://www.anthropic.com/index/evaluating-ai-systems

[25] 大槻繁 思考の技法のあゆみとゆくえ，知働化研究会誌 Vol.3 （超マシーン）

[26] 濵勝巳 セル組織，知働化研究会誌 Vol.3

[27] Society of Automotive Engineers SAE International. September 2016,
taxonomy and de nitions for terms related to driving automation systems for
on-road motor vehicles. http://standards.sae.org/j3016 201609/

[28] OpenAI 『Practices for Governing Agentic AI Systems』

https://openai.com/research/practices-for-governing-agentic-ai-systems

[29] 自己決定／自律」および「自己決定権」についての基礎的考察
https://www.r-gscefs.jp/pdf/ce01/hy01.pdf

[30] World Models for Autonomous Driving An Initial Survey
https://arxiv.org/abs/2403.02622

[31] Penrose The Road to Reality : A Complete Guide to the Laws of the Universe

[32] Wolfram A New Kind of Science https://www.wolframscience.com/nks/

[33] A Closer Look at Large Language Models Emergent Abilities
https://yaofu.notion.site/A-Closer-Look-at-Large-Language-Models-Emergent-Abilities-493876b55df5479d80686f68a1abd72f

[34] Language Models of Code are Few-Shot Commonsense Learners
https://arxiv.org/pdf/2210.07128.pdf

[35] Chain of Thought https://www.promptingguide.ai/jp/techniques/cot

[36] Brain-Inspired Two-Stage Approach: Enhancing Mathematical Reasoning by Imitating Human Thought Processes https://arxiv.org/abs/2403.00800

[37] Executing Natural Language-Described Algorithms with Large Language Models: An Investigation https://arxiv.org/abs/2403.00795

■