OpenAI の AGI 開発 5 段階フレームワーク：LLM 層とアプリケーション層の役割と進化

Gemini Deep Researh

I. OpenAI の AGI 開発フレームワーク序論

A. 汎用人工知能（AGI）の定義

汎用人工知能（AGI）は、人間と同様の認知能力を備え、人間が実行可能なあらゆる知的タスクを理解し、学習し、実行できる AI システムとして構想されている[1]。これは、特定のタスクに特化した現在の「狭い AI」とは対照的である。AGI は現時点では理論的な概念であり、AI 分野における主要な研究目標の一つと位置づけられている[3]。

B. OpenAI の 5 段階分類システム

OpenAI は、AGI 達成に向けた AI 能力の進捗を分類し追跡するための 5 段階のフレームワークを開発した[4]。このフレームワークは、技術アナリスト、経営幹部、研究者が AI の進化とその潜在的な影響を理解するための一助となることを目的としている[4]。留意すべき点として、OpenAI のフレームワークは AGI への道筋を定義する試みの一つであり、Google DeepMind のような他の主要研究機関も同様の、しかし異なるレベル定義を提案している[2]。このようなフレームワークの多様性は、AGI の正確なマイルストーンや達成への最適経路に関して、学術界および産業界で活発な議論が継続していることを示唆している。

複数の AGI フレームワーク（OpenAI のものや Google DeepMind のものなど）が同時に存在し、それぞれが推進されているという事実は、AGI への道筋がまだ普遍的に成文化されたものではないことを示唆している。これは、AGI 自体が活発に定義されつつある概念であり、主要な研究主体が異なる中間能力に最適化していたり、進捗に対して異なる測定基準を使用していたりする可能性を示している。例えば、OpenAI の機能的役割に基づくレベル（チャットボット、推論者）と、Google DeepMind のパーセンタイルに基づく能力レベル（有能、専門家）との間のレベル記述の違い[2]は、AGI の「ゴールライン」が異なって概念化されているか、あるいは道中の重要な中間マイルストーンが主要な研究室によって異なる重み付けをされていることを示唆している。この相違は、AI 能力のより広範で多様な探求を促進するかもしれないが、進捗の直接的な比較を複雑にする可能性もある。

OpenAI のフレームワークが、そのレベルを機能的な記述（チャットボット、推論者、エージェント、イノベーター、組織）で構成していることは、AI の進歩を認識可能な

タスクや役割を実行する実用的な有用性と本質的に結びつけているように見える。これは、より抽象的な認知ベンチマークを優先する可能性のあるフレームワークとは対照的である。この方向性は、AI を現実世界の経済的および社会的機能にますます統合し、価値あるものにするという OpenAI の戦略的重点を反映している可能性がある。レベルの名称自体[4]は、AI が実行できる作業や相互作用の種類を直接想起させる。このアプローチは、AGI 開発が AI の経済的に価値のある作業を遂行する能力によってベンチマークされるという OpenAI の根底にある哲学を示唆しており、レベル 5 の AGI がそのようなタスクで人間の能力を凌駕するという記述で明確に示されている[6]。この実用的で応用中心の視点は、OpenAI の研究優先順位と AGI に向けた「進捗」の定義を形成している可能性が高い。

AGI への多段階の道筋を定義する行為そのものは、たとえそれが「試案」[2]であったとしても、研究課題を構造化し、（内部的にも外部的にも）期待を設定し、潜在的にリソースと人材を引き付けるのに役立つ。しかし、[38]および[2]で批判されているように、レベル間の質的な違いと規模が厳密に定義され、一貫して適用されない場合、そのようなフレームワークは「マーケティング」または過度の単純化と見なされる可能性もある。明確なロードマップと AGI 固有の複雑さとの間のこの緊張関係は、重要な根底にあるテーマである。

C. 本報告書の目的と範囲

本報告書は、OpenAI の AGI 開発 5 段階レベルについて詳細かつ分析的な解説を行うことを目的とする。各開発段階における大規模言語モデル（LLM）層とアプリケーション層の明確かつ進化する役割、機能、相互作用の分析に重点を置く。

II. 高度 AI システムの基盤層

A. 大規模言語モデル（LLM）層

コア能力

LLM は AI の特殊なサブセットであり、特に深層学習モデルとして、人間の言語を驚くほどの流暢さで処理、理解、生成することに長けているように設計されている[8]。その開発は、膨大な量のテキストデータ（そしてますますマルチモーダルなデータ）でのトレーニングに依存している[8]。

主要機能

LLM の機能レパートリーは広範であり、高度なテキスト要約、多言語翻訳、ニュアンスのある質疑応答、多目的なコンテンツ生成（テキスト、画像、音楽、ソフトウェアコ

ードを含む）、詳細な意味解析、複雑なパターン認識、高度なデータ分析などのタスクを網羅している[8]。

アーキテクチャ基盤

トランスフォーマーアーキテクチャは、その自己注意メカニズムとともに、現代のLLM 設計の基礎であり、シーケンシャルデータを効果的に処理し、入力データのさまざまな部分の重要性を重み付けすることを可能にしている[10]。

「インテリジェンスエンジン」としての役割

LLM 層は AI システム内の中核的な認知エンジンとして機能する。自然言語理解、推論、知識表現、およびより複雑な AI アプリケーションと機能が構築される生成タスクの基本的な能力を提供する[9]。

B. アプリケーション層

定義と機能

AI システムの文脈において、アプリケーション層は、AI 駆動の機能が実装され、エンドユーザーまたは他のソフトウェアシステムに直接アクセス可能になるソフトウェアスタックの層を表す[12]。この層には通常、人間との対話のためのユーザーインターフェース（UI）と、プログラムによるアクセスおよびシステム統合のためのアプリケーションプログラミングインターフェース（API）が含まれる[12]。

AI 能力提供における役割

アプリケーション層は、AI モデルの抽象的な能力を具体的で利用可能なツールやサービスに変換するために不可欠である。これは、事前にトレーニングされた AI モデルまたはリアルタイムの機械学習アルゴリズムを統合し、アプリケーション固有のロジックを管理し、ユーザー中心の視点からデータインタラクションを処理し、AI が生成した出力と洞察を理解しやすい方法で提示することによって達成される[12]。

具体例

顕著な例としては、音声アシスタント（Siri、Alexa など）のような会話型 AI、自動運転車内の高度なソフトウェア、電子商取引における AI 駆動の推奨エンジン、医療診断を支援する AI ツールなどが挙げられる[12]。

C. 共生的かつ進化する関係

相互依存性

基本的な共生関係が存在する。LLM 層は中核的な「知能」（例：言語理解、推論）を提供し、アプリケーション層は特定の問題やタスクにこの知能を効果的に適用するための必要な「インターフェース、コンテキスト、ツール」を提供する。どちらの層も、単独では高度な AI 機能を実現できない。

共進化

LLM とアプリケーション層の開発は、密接に連携した共進化プロセスである。LLM がより強力で多用途になるにつれて、アプリケーション層もこれらの新しい能力を活用するために進化し、より高度で自律的で機能豊富なアプリケーションの作成につながる必要がある。逆に、複雑なアプリケーション要件（例：信頼性の高いツール使用やマルチエージェントオーケストレーションの必要性）によってもたらされる要求と課題は、LLM 開発における革新と改良の重要な推進力となる。この動的な相互作用は、OpenAI の AGI レベルの進行を通じて明らかである。

アーキテクチャパラダイム

LLM 層とアプリケーション層間のインターフェースと統合は、さまざまな方法で設計できる。AI システムが複雑さを増すにつれて、特に高度な AGI レベルに向けて、モジュール式の分離されたアーキテクチャへの重点が高まっている。例えば、3 層分離アーキテクチャ（アプリケーション層、プロトコル層、ハードウェア層で構成）[14] は、LLM 駆動アプリケーションにおけるモジュール性、相互運用性、スケーラビリティを向上させるために提案されている。

「LLM 層」に帰属する機能範囲は静的なものではなく、AGI 開発スペクトル全体で劇的に拡大する。主に自然言語処理に焦点を当てたレベル 1 の LLM は、革新能力が求められるレベル 4 で期待される高度な認知エンジン（しばしば LLM またはその後継機と呼ばれる）とは質的に異なる。これは、「LLM」が高度 AI の中核的な推論、生成、そして潜在的には創造的なエンジンのためのより広範な用語に進化していることを意味する。LLM の初期の定義 [8] は言語タスクを強調している。しかし、OpenAI の AGI フレームワーク [4] は、「博士号レベルの問題解決」（レベル 2）や「イノベーションの生成」（レベル 4）を、これらの高度なモデルから中核的な知能を引き出す AI システムに帰属させている。これは、「LLM 層」が単なる言語処理を超えて、より広範な一般認知機能を包含することが期待されていることを示している。「基盤モデル」[10] という記述の使用も、この拡大する基本的な役割の概念を支持している。

アプリケーション層の進化は、単なる UI/UX の審美的な強化だけではない。内部ロジックの複雑さが大幅に増し、高度なタスクオーケストレーション、動的なツール統合と使用、マルチエージェント調整フレームワーク、および多様な外部システム、データス

トリーム、API とのシームレスな統合を網羅するようになる。これにより、アプリケーション層自体が AI 研究およびエンジニアリングの重要な領域として位置づけられる。レベル 1 では、アプリケーション層は比較的単純なチャットインターフェースかもしれない[12]。レベル 3（エージェント）までには、長期間にわたる自律的なタスク実行をサポートする必要があり[1]、スケジューリング、状態管理、外部 API やツールとの堅牢な相互作用における高度な能力が必要となる[14]。レベル 5（組織）では、アプリケーション層は企業全体の複雑で相互接続されたワークフローを管理および実行する必要がある[4]。この軌跡は、アプリケーション層が LLM 出力の単純な導管から、高度な AI 能力を実現するために不可欠な複雑なオーケストレーションおよび実行環境へと変貌することを示している。

安全性、セキュリティ、公平性、透明性を包含する「信頼できる AI」[10] への要求の高まりは、LLM 層とアプリケーション層間のますます緊密で共同設計された統合を必要とするだろう。効果的な安全プロトコル、バイアス緩和戦略、セキュリティ対策は、別々の層への分離されたアドオンとして実装することはできず、それらの相互作用の構造に織り込まれなければならない。プロンプトインジェクション[16] や安全でない出力処理[16] などのセキュリティ脆弱性は、ユーザー（アプリケーション層経由）が LLM と対話するインターフェースで本質的に発生する。アルゴリズムバイアス[22] などの倫理的懸念は、しばしば LLM トレーニングデータに起因するが、アプリケーションの出力を通じてユーザーに顕在化し影響を与える。したがって、効果的な緩和には相乗的なアプローチが必要である。アプリケーション層は入力を積極的にサニタイズし、出力を検証し、潜在的にコンテキストに応じた保護手段を実装する必要があり、一方、LLM 層はより大きな固有の堅牢性、整合性、操作への耐性のために設計される必要がある。これは、安全でセキュアで倫理的な AI システムを構築するための、深く絡み合った全体的な開発プロセスを示している。

III. OpenAI の AGI5 段階レベルの詳細分析

表 1：OpenAI の AGI レベルの概要：LLM 層とアプリケーション層の特性と進化

レベル番号と名称	このレベルでの主要 AI 能力	主要 LLM 層の機能と進化	主要アプリケーション層の機能と進化	主要な技術的ハードル	新たな倫理的懸念
レベル 1：チャットボッ	会話能力、基本的な指示理	基本的な NLU/NLG、	チャット UI、基本的な	ハルシネーション、一貫	バイアス、プライバシー侵

ト（会話型 AI）	解	短期コンテキスト維持、暗黙的な情報検索	API 統合、セッション管理	性、限定的な推論、知識のカットオフ	害、誤用、表面性
レベル 2：推論者（人間レベルの問題解決）	博士号レベルの問題解決能力（ツールなし）、論理的思考、深い文脈理解	高度な論理推論、深い文脈理解、仮説生成（初歩的）、ハルシネーションの低減、マルチモーダル推論（萌芽的）	高度な分析ツール、研究支援、診断支援、複雑な Q&A システム、限定的なツール使用のオーケストレーション	推論の一般化可能性、説明可能性（XAI）、推論の効率性、決定論対確率論	誤った推論に対する説明責任、過度の依存、アクセスと公平性、人間レベルの曖昧さ、道徳的推論能力
レベル 3：エージェント（自律型 AI）	数日間にわたる自律的なタスク実行、意思決定、適応	長期計画と目標維持、ツール使用と API 連携、対話からの動的適応と学習、永続的メモリと状態管理、不確実性下での意思決定	AI エージェント展開・管理プラットフォーム、高度なパーソナルアシスタント、自律型ワークフローシステム、ロボット制御	信頼性と堅牢性、安全性と制御（アラインメント）、スケーラビリティ、リソース管理、ツール誤用	説明責任、プライバシー侵害、セキュリティリスク（過剰な権限付与）、雇用喪失、人間の監督と制御の喪失
レベル 4：イノベーター（創造的 AI）	新しい発明・アイデアの自律的創出、独創的思考	抽象的概念化と類推、創造的仮説生成、分野横断的知識統合、制約のない探求、新規性と有用性の評価	AI 駆動 R&D プラットフォーム、自動科学発見ツール、創造的コンテンツ生成スイート、創薬プラットフォーム	「新規性」「創造性」の定義と評価、自律的な問題再定義、実現可能性の担保、創造性の誘導と制御、イノベーションのためのマルチモーダル統合	知的財産権、イノベーションにおけるバイアス、予期せぬ結果、人間のイノベーターの役割、AI 生成科学の信頼性、有害なイノベーションに対する説明責任
レベル 5：組	組織全体の機	包括的システ	完全自律型	極度の複雑性	実存的リス

織（AGI/ASI プロキシ）	能遂行、複雑なワークフロー管理、戦略的思考、自己改善	ム理解、戦略的推論と長期計画、複数 AI エージェント/サブシステムの調整、大規模な倫理的意思決定、継続的な自己改善と適応	AI 経営企業/組織単位、AI CEO、AI マネージャー、AI 駆動運用プラットフォーム	管理、堅牢性と回復力、大規模な価値整合（スーパーアラインメント）、解釈可能性とデバッグ	ク、権力集中、社会的混乱（大量失業）、ガバナンスと制御、公平性と公正性、セキュリティ（組織レベルでの脆弱性）

A. レベル 1：チャットボット（会話型 AI）

1. 能力と特性の定義

この基礎レベルの AI システムは、会話タスクにおける習熟度によって特徴づけられる。自然言語入力を理解し、人間らしい応答を生成することで首尾一貫した対話を行い、明確なユーザー指示に基づいて簡単なタスクを実行できる[4]。顕著な例としては、カスタマーサービスチャットボットや、OpenAI の ChatGPT、Anthropic の Claude、Google の Gemini などの広く利用されている仮想アシスタントが挙げられる[1]。このレベルは、広くアクセス可能な AI の現状を表している。GPT-4o のような高度なモデルでさえ、一般的にレベル 1 に位置づけられるが、レベル 2 に関連する初期の能力をますます示し始めている[4]。

2. LLM 層の役割と進化

クリティカルな LLM 機能

- **自然言語理解（NLU）と自然言語生成（NLG）：** これらは、ユーザーの問い合わせの意味と意図を理解し、流暢で首尾一貫し、文脈的に適切な応答を生成するために最も重要である[8]。
- **基本的なコンテキスト管理：** LLM は会話の流れを維持し、対話の継続性を確保するために最近のやり取りを記憶する必要がある。
- **情報検索（暗黙的）：** LLM は、事実に基づいた質問に答えたり、関連情報を提供したりするために、トレーニングデータに埋め込まれた膨大な知識にアクセスし、利用する。

LLM アーキテクチャ/トレーニング

レベル 1 の LLM は通常、トランスフォーマーアーキテクチャに基づいている。文法、意味論、事実知識、会話パターンを学習するために、大規模で多様なテキストコーパス（および時には他のモダリティ）で事前トレーニングされる[10]。この段階での主要な進化的焦点は、会話の流暢性の向上、より長い対話にわたる一貫性の改善、および無意味または無関係な出力の生成の削減にある。

3. アプリケーション層の役割と進化

アプリケーションタイプ

レベル 1 のアプリケーション層は、主にチャットインターフェース（テキストベースまたは音声対応）、さまざまなデバイスやサービスに統合された仮想アシスタントプラットフォーム、およびカスタマーサポート自動化のための組み込みツールとして現れる[12]。

LLM の活用

アプリケーション層は仲介役として機能する。プロンプト（テキストまたは音声）を入力するためのユーザーインターフェースを提供し、これらのプロンプトを処理のために LLM に送信し、その後、LLM が生成した応答を受信してユーザーに表示する[12]。また、セッション管理のための単純なバックエンドロジックや、特定の狭く定義されたタスク（例：外部 API を介した現在の天気情報の取得）のための初歩的な API 統合を組み込む場合もある。

複雑性

このレベルのアプリケーションロジックは比較的単純であり、主な重点は、スムーズなユーザーと LLM の会話を促進し、情報を明確でアクセスしやすい方法で提示することにある。

4. 主要な技術的ハードルと新たな倫理的考慮事項

技術的ハードル

- **ハルシネーション：** LLM がもっともらしいが事実と異なる、無関係な、または完全に無意味な応答を生成する傾向は、重大な課題である[6]。
- **一貫性：** 長時間または複雑な会話全体を通じて、一貫したペルソナ、トーン、および事実の正確性を維持することは依然として困難である。
- **限定的な推論：** レベル 1 の LLM は一般的に、直接的な情報検索や単純な推論を超える複雑な多段階推論に苦労する。
- **知識のカットオフ：** LLM の知識は最後のトレーニング日までのものであり、拡張

されない限り、ごく最近の出来事については認識していない。

倫理的考慮事項

- **バイアス：** LLM は、広範なトレーニングデータに存在する社会的バイアスを不注意に反映し、さらに増幅させる可能性があり、不公平、差別的、またはステレオタイプを永続させる出力につながる可能性がある [17]。
- **プライバシー：** LLM がトレーニング中に遭遇した機密情報や個人情報を記憶し、漏洩するリスクがある [22]。
- **誤用：** レベル 1 の AI の機能は、スパムの生成、説得力のあるフィッシングメールの作成、誤情報やプロパガンダの拡散など、悪意のある目的で悪用される可能性がある [22]。
- **表面性と過度の依存：** AI が生成した応答は、深み、ニュアンス、または真の理解を欠く場合がある。特に教育的または意思決定の文脈でそのような応答に過度に依存すると、批判的思考と学習が妨げられる可能性がある [26]。

レベル 1 における主要な開発の推進力は、会話の流暢性を達成することを超えて、生成されるテキストが検証可能な事実に裏打ちされ、文脈的に適切であり、安全ガイドラインを遵守することを保証することにますます焦点が当てられている。これは、LLM が単なる統計的パターン複製を超えて、知識検証と倫理的境界認識における初期の能力を備える必要があることを意味する。ハルシネーション [6] やバイアス [22] といった根強い問題は、LLM が純粋に根拠のない生成モデルである場合の限界を直接露呈している。レベル 1 では、アプリケーション層はしばしば LLM 応答の直接的な導管として機能するため、LLM の欠陥はすぐにユーザーに露呈する。この現実は、LLM 開発者に対し、単に言語の流暢性を最適化するだけでなく、事実性、バイアスの低減、安全性の向上を強化するメカニズムを組み込むよう大きな圧力をかけている。

レベル 1 の AI 用アプリケーション層は、技術的には単純に見えるかもしれないが（例：チャットウィンドウ）、その設計は、AI システムに対するユーザーの認識を形成し、期待を管理し、信頼（または不信）を育む上で、不均衡に重要な役割を果たす。思慮深く設計されたインターフェースは、チャットボットの認識される有用性を高め、LLM の限界に対するユーザーの不満を軽減することができる。[12] および [13] の記述は、ユーザーインタラクションにおけるアプリケーション層の極めて重要な役割を強調している。基礎となる LLM 機能が同一であっても、情報の提示方法、インタラクションの容易さ、AI の限界の明確さ、およびアプリケーションによる会話状態管理の有効性は、ユーザーエクスペリエンスに大きな影響を与える可能性がある。これは、AI 開発のこの初期段階でさえ、成功した AI の採用と有用性のためには、ヒューマン・コンピュータ・インタラクション（HCI）とユーザー中心設計の原則が不可欠であることを強

調している。

レベル 1 の AI の「会話的」な性質は諸刃の剣である。アクセシビリティとユーザーエンゲージメントを劇的に向上させる一方で、ユーザーがシステムを擬人化する場合、微妙な操作や AI の真の理解の過大評価に対する潜在的な脆弱性を同時に生み出す。チャットボットが人間のような対話を模倣する能力[4]は、ユーザーが AI に実際に備わっている以上の知性、意識、または理解を容易に帰属させることにつながる可能性がある。これはユーザーエンゲージメントには有益であるが、ユーザーが特に機密性の高いトピックに関して、独立した検証なしに AI が生成したアドバイス、事実の主張、さらには説得力のある議論を無批判に受け入れる場合、問題となる。このリスクは、偽情報のための誤用[22]や応答が表面的である可能性[26]に関する懸念によって増幅される。

B. レベル 2：推論者（人間レベルの問題解決）

1. 能力と特性の定義

レベル 2 の AI システム、「推論者」と呼ばれるものは、博士号を持つ個人に匹敵する人間レベルの問題解決能力を示す能力によって定義されるが、専門的な外部ツールやトレーニングを超えたリアルタイムの情報アクセスなしに動作する[4]。これらの AI は、論理的思考、多段階推論、および多様なドメインにわたる文脈の深い理解を必要とする複雑なタスクに取り組むことが期待されている[4]。OpenAI の「o1」モデルがこのレベルに達したか、またはこのレベルを代表するものであると示唆されている[1]。この段階での重要な進歩は、AI ハルシネーションの大幅な削減であり、より正確で信頼性が高く、検証可能な応答の提供につながる[6]。レベル 2 の AI のパフォーマンスは、熟練した成人人間の 50％を、彼らが熟達しているタスクにおいて上回るものとしてベンチマークされている[1]。

2. LLM 層の役割と進化

クリティカルな LLM 機能

- **高度な論理的および推論的推論：** LLM は、複雑な多段階の論理的演繹、帰納、アブダクション推論を実行し、抽象的な概念を効果的に処理できなければならない[27]。
- **深い文脈理解：** これには、多面的な問題内の微妙なニュアンス、暗黙の情報、因果関係、および複雑な相互依存関係を理解することが含まれる。
- **仮説生成と評価（初歩的から中級）：** もっともらしい仮説または潜在的な解決策を策定し、利用可能な情報に基づいてそれらの妥当性または可能性を評価する能力[27]。
- **強化された事実の正確性とハルシネーションの低減：** このレベルの LLM は、内部

的な一貫性チェック、事実検証（トレーニングデータに対する）、自己修正、および自身の出力の信頼性評価のための改善されたメカニズムを組み込む必要がある[6]。

- **出現するマルチモーダル推論：** OpenAI の o3 のような最近の高度なモデルは、テキストや画像など、複数のモダリティからの情報を統合し、それらを単に個別に処理するだけでなく、推論する能力を示している[28]。

LLM アーキテクチャ/トレーニング

レベル 2 への進化には、LLM アーキテクチャの大幅な進歩、または推論能力を強化するために特別に設計された新しいファインチューニング方法論の開発が伴う可能性が高い（例：o3 について説明されているように、「応答する前により長く考えるようにトレーニングされた」モデル[29]）。トレーニングデータセットは、より複雑な問題セット、科学文献、論理パズル、数学的証明、および構造化された知識ベースで強化され、高度な推論スキルを育成することが期待される。高度な人間/AI フィードバックからの強化学習（RLHF/RLAIF）や憲法 AI 原則などの技術は、単なる流暢さやもっともらしさよりも、健全な推論、事実の正確性、および真実の応答を優先するように改良される可能性が高い。

3. アプリケーション層の役割と進化

アプリケーションタイプ

アプリケーション層は、より高度な分析ツール、AI 駆動の研究アシスタント、高度な診断支援（例：医療または工学）、および深い推論を必要とする複雑で多面的な問い合わせに対応できる複雑な質疑応答システムを可能にする。

LLM の活用

アプリケーション層の役割は、単純なプロンプトの受け渡しを超える。複雑な問題を LLM 処理に適した形式に構造化し、多段階の推論対話を管理し、複雑な問い合わせをサブクエリに分解し、潜在的にキュレーションされた外部知識ベースまたは検証サービスと統合して LLM の推論プロセスを増強および検証することが含まれる場合がある。

複雑性

アプリケーションロジックはかなり複雑になる。より構造化された複雑な入力と出力を処理する必要があり、特に重要な意思決定のためには、LLM の推論ステップをユーザーに視覚化または説明する機能を提供する必要があるかもしれない。LLM がどの（限定的な）ツールを使用するかを決定する、LLM 主導のツール使用の初期の形態も、アプリケーション層によって調整されて現れ始める可能性がある[28]。

4. 主要な技術的ハードルと新たな倫理的考慮事項

技術的ハードル

- **推論の一般化可能性：** AI の推論能力が、特定のベンチマークタスクやトレーニングデータ分布に狭く限定されるのではなく、広範な多様なドメインや新しい問題タイプにわたって堅牢で移転可能であることを保証することが大きな課題である。
- **説明可能性（XAI）：** 推論がより複雑になるにつれて、LLM が特定の解決策にどのように到達したかを理解し明確に説明する能力は、信頼を構築し、デバッグを可能にし、説明責任を確保するために不可欠になる。
- **推論のスケーラビリティと効率性：** 深い多段階推論を実行することは計算集約的である可能性がある。これを効率的かつ大規模に実行できることを保証することが重要である。
- **決定論対確率論：** LLM 固有の確率論的（確率的）な性質は、多くの科学的および企業的文脈で一般的な、正確で再現可能で検証可能な推論を必要とするアプリケーションにとって重大な問題となる可能性がある [30]。

倫理的考慮事項

- **誤った推論に対する説明責任：** 「推論者」AI が欠陥のあるアドバイスや誤った解決策を提供し、それが現実世界で否定的な結果（例：医療診断や財務計画における）をもたらした場合、責任と説明責任を決定することは複雑な倫理的および法的問題となる [31]。
- **過度の依存とスキルの低下：** ユーザーが AI の問題解決能力に過度に依存するようになり、人間の批判的思考とドメイン固有の専門知識が低下するリスクがある。
- **アクセス、公平性、推論におけるバイアス：** 強力な推論 AI ツールが公平にアクセス可能であり、問題解決アプローチにおいて既存の社会的バイアスを永続させたり増幅させたりしないようにすることが不可欠である。偏った推論は差別的な結果につながる可能性がある。
- **「人間レベル」ベンチマークの曖昧さ：** 「博士号レベル」の推論というベンチマークは定義が曖昧であり、分野によって大きく異なり、人間の専門家の推論のすべての側面を捉えていない可能性がある（[2] の批判）。
- **AI の道徳的推論能力：** AI システムがより高度な推論を発達させるにつれて、道徳的推論能力、倫理的枠組みの源泉と性質、およびこれらが多様な人間の価値観とどのように整合するかについて、深遠な問題が生じる [25]。

「推論者」への移行は、極めて重要なパラダイムシフトを意味する。LLM は、主に洗練された言語処理装置から、多目的な認知ツールへと進化している。これには、統計的パターンマッチングや流暢なテキスト生成を超える LLM 能力の根本的な強化が必要で

あり、記号的推論、論理的推論、因果的理解を機能的に近似するメカニズムを組み込む必要がある。レベル 1 の AI は主に会話能力によって特徴付けられる[4]。対照的に、レベル 2 は「人間レベルの問題解決」と「論理的思考」を要求する[4]。これは、基礎となる LLM が、学習した言語パターンに基づいて次のトークンを予測するだけでなく、知識と概念の内部表現を操作して解決策を導き出す必要があることを意味し、その認知機能における質的な飛躍を表している。このレベルでのハルシネーションの軽減と事実の正確性の向上への強い重点[6]は、内部的な一貫性チェック、論理的検証、およびある程度の「真実性」の要件をさらに示しており、これらはすべて堅牢な推論の不可欠な側面である。

「推論者」向けに設計されたアプリケーション層は、より構造化された対話的なエンゲージメントモードをサポートするために大幅に進化する必要がある可能性が高い。また、「説明可能性」と透明性を高める機能を組み込む必要もあるだろう。複雑な問題の場合、ユーザーは解決策だけでなく、特に利害関係が大きい場合には、基礎となる推論プロセスをますます理解する必要があるだろう。AI システムが「博士号レベル」[4]で動作している場合、それが生成する出力は重要な意思決定に情報を提供する可能性が高い。その結果、ユーザー（例：科学者、エンジニア、医療専門家）は AI の出力に対して高度な信頼を必要とするだろう。そのような信頼は、透明性と AI の推論を精査する能力によって大幅に育まれる。したがって、アプリケーション層は、単純なチャットインターフェースを超えて進化し、最終的な解決策だけでなく、「思考の連鎖」、裏付けとなる証拠、または LLM の推論プロセスに関連する信頼スコアも提示する必要があるかもしれない。これには、より高度な UI/UX 設計と潜在的に新しい対話パラダイムが必要となる。

堅牢で一般化可能なレベル 2 の推論能力を達成することは、多くの分野で科学的発見、工学的革新、複雑な意思決定を劇的に加速させる可能性がある。しかし、この力は、AI の推論に欠陥がある場合、微妙または明白でない方法で偏っている場合、または新しい状況に直面したときに脆弱である場合、重大なリスクももたらす。「博士号レベル」という主張自体、専門知識の幅と深さ、および AI の出力に対する過信の可能性に関して慎重な精査が必要である。複雑な問題を解決する能力[4]は、レベル 2 の AI を非常に強力なツールにする。しかし、[2]で強調されている批判（「すべてのドメインで博士号？」）は、表面的または狭く専門化された専門知識の可能性に関する重要な点を提起している。AI の推論が脆弱であるか、トレーニングデータに過剰適合しているか、または真の一般化を欠いている場合、それを新しい、利害関係の大きい問題に適用することは危険である可能性がある。エラーに対する説明責任[31]に関する倫理的考慮事項は、AI の推論プロセスが不透明であるか、またはその限界が十分に理解されていない場合、はるかに深刻になる。したがって、アプリケーション層は、AI の出力の厳

密な検証と人間の専門家による批判的な評価を促進する必要がある。

C. レベル 3：エージェント（自律型 AI）

1. 能力と特性の定義

レベル 3 の AI システム、通称「エージェント」は、顕著な自律性を示す。ユーザーに代わって長期間、場合によっては数日間にわたりタスクを実行し、意思決定を行う能力を有する[1]。これらのシステムは、受動的な「コパイロット」やアシスタントから、目標を積極的に追求できる能動的な「タスクマネージャー」への移行を示す[4]。例としては、複雑なスケジュールの管理、メール対応の自律処理、高度にパーソナライズされた推奨の提供、さらにはソフトウェアエンジニアリングの側面など、より複雑なタスクの実行が挙げられる（例：「Devin」というシステムはエージェント型 AI の例として言及されているが、OpenAI の哲学では、このような高度なエージェンシーの前に堅牢な推論が必要であると示唆されている）[6]。重要な特徴は、絶え間ない人間の介入を必要とせずに、変化する状況や新しい情報に応じて独立して意思決定を行い、計画や行動を適応させる能力である[1]。レベル 3 の AI は、設計されたタスクにおいて熟練した成人人間の 90%を上回るパフォーマンスを発揮するとベンチマークされている[1]。OpenAI の戦略的ロードマップは、堅牢な推論能力（レベル 2）が効果的で信頼性が高く安全なエージェントを開発するための基礎的な前提条件であることを強調している[6]。

2. LLM 層の役割と進化

クリティカルな LLM 機能

- **長期計画と目標分解：** LLM は、高レベルの目標を理解し、それらを実行可能なサブタスクに分解し、長期間にわたって一貫した計画を維持できなければならない。
- **熟練したツール使用と API 連携：** 中核機能は、外部ツール、ソフトウェアライブラリ、API をインテリジェントに選択、呼び出し、結果を解釈して情報を収集し、外部システムと対話し、デジタル環境および潜在的に物理環境でアクションを実行する能力である[29]。
- **動的適応と対話からの学習：** エージェントは、新しい情報、受け取ったフィードバック（ユーザーまたは環境から）、または状況の予期せぬ変化に基づいて、計画と行動を修正できなければならない。これは、何らかの形の継続的または対話的な学習能力を意味する。
- **永続的メモリと状態管理：** タスクの進捗、ユーザーの好み、環境の状態、および関連するコンテキスト情報の堅牢で永続的な理解を長期間維持することは、一貫した長期的なエージェンシーにとって不可欠である。
- **不確実性と曖昧さの下での意思決定：** エージェントはしばしば不完全または曖昧

な情報を持つ環境で動作するため、LLM は合理的な選択を行い、不確実性を管理する必要がある。

LLM アーキテクチャ/トレーニング

このレベルのモデルは、高度な強化学習技術を使用して、おそらく対話的なシミュレートされた環境内、または成功したタスク完了とツール使用に報酬を与える特殊なエージェントトレーニングフレームワークを通じてトレーニングされる可能性が高い。トレーニング中の重要な重点は、多様なツールを効果的、安全、かつ確実に使用することを学習することである（例：ツールの誤用に対する保護手段を実装するための LlamaFirewall のようなフレームワーク [16]）。アーキテクチャ的には、レベル 3 のエージェントは、中央の LLM が特定の機能に合わせて調整された特殊なサブモジュール、より小さな LLM、または他の AI コンポーネントを調整するモジュラー設計を含む場合がある。

3. アプリケーション層の役割と進化

アプリケーションタイプ

アプリケーション層は、AI エージェントの展開、管理、対話のための高度なプラットフォームに進化する。これには、重要なエージェンシーを持つ高度にパーソナライズされたデジタルアシスタント、自律的なワークフロー自動化システム、および潜在的にロボットエージェントの制御システムが含まれる [30]。

LLM の活用

アプリケーション層は、エージェントに不可欠なインフラストラクチャを提供する。これには、LLM 駆動エージェントに必要なツール、データストリーム、および実行環境への安全なアクセスを許可することが含まれる。権限の管理、安全性とパフォーマンスのためのエージェント活動の継続的な監視、およびユーザーが複雑なタスクを委任し、高レベルの目標を定義し、進捗更新を受信したり介入を要求したりするための直感的なインターフェースの提供を担当する。

複雑性

レベル 3 のアプリケーション層は非常に複雑になる。ツール統合のための堅牢で安全な API、意図しないシステム全体への影響を防ぐためのサンドボックス化された実行環境、監査とデバッグのための高度な監視およびロギング機能、および長期間実行される自律的なタスクを管理するために設計されたユーザーインターフェースを備えている必要がある。3 層分離モデル（アプリケーション、プロトコル、ハードウェア）のアーキテクチャ原則は、マルチエージェントオーケストレーションを管理し、クロスプラット

フォームの相互運用性を確保し、システムの安定性を維持するために非常に関連性が高くなる[14]。アプリケーション層は本質的に、これらの自律エージェントの「オペレーティングシステム」または「ランタイム環境」として機能する。

4. 主要な技術的ハードルと新たな倫理的考慮事項

技術的ハードル

- **信頼性と堅牢性：** エージェントが予期せぬ出来事やノイズの多い環境に直面しても、長期間にわたって複雑なタスクを正しく、一貫して、安全に実行できることを保証することは、最も重要な課題である。
- **安全性、制御、アラインメント：** 自律エージェントが有害な、意図しない、または誤った行動をとることを防ぐこと（しばしば「アラインメント問題」と呼ばれる）は、重大な懸念事項である。人間の監督、介入（「キルスイッチ」）、および堅牢な安全プロトコルのメカニズムは不可欠である[30]。
- **スケーラビリティ：** 潜在的に多数の自律エージェントを効率的に管理および調整できるシステムを開発すること。
- **リソース管理：** 長時間実行されるエージェントが計算リソース（処理、メモリ、ネットワーク帯域幅）を効率的に利用することを保証すること。
- **関数呼び出しハルシネーションとツール誤用：** エージェントが指示や文脈を誤解し、ツールの誤った選択や使用につながり、エラーや意図しない結果を引き起こす可能性がある[37]。

倫理的考慮事項

- **説明責任と責任：** 自律エージェントが危害を引き起こしたり、重大なエラーを犯したり、倫理規範に違反したりした場合に、明確な説明責任と法的責任の所在を確立することはますます困難になる[36]。
- **プライバシー：** 自律エージェントは、タスクを効果的に実行するために、広範な個人情報、機密情報、または機密データへのアクセスを必要とする場合があり、重大なプライバシー懸念を引き起こす[25]。
- **セキュリティリスク：** AI エージェント自体が高度な攻撃の標的になったり、侵害された場合に悪意を持って使用されたりする可能性がある（例：過剰な権限付与につながるプロンプトインジェクション[16]）。
- **雇用喪失：** 複雑なタスクを実行する自律エージェントは、大幅な雇用喪失につながる可能性がある[37]。
- **人間の監督と制御の喪失：** エージェントが人間の意図と一致しない目標を追求するリスク[36]。行動シーケンスに基づく規制が提案されている[36]。

「エージェント」への移行は、AI が情報処理者/推論者からデジタルまたは物理世界の行為者へと移行することを意味する。これにより、潜在的な影響（肯定的および否定的の両方）と安全性およびアラインメントの複雑さが劇的に増大する。レベル 1 および 2 は主に理解と応答に関するものである。レベル 3 は、ユーザーに代わってタスクを実行し意思決定を行うこと [4]、つまり行うことに関するものである。この積極的な役割は、AI がもはや人間がその後行動する情報を提供するだけでなく、それらの行動を自ら行うことを意味する。これにより、偽情報（レベル 1/2）から潜在的に有害な自律的行動（レベル 3）へとリスクがエスカレートし、[36] および [37] の安全懸念で強調されている。

エージェント向けアプリケーション層は、AI システムの安全性とガバナンスの重要な構成要素となる。それはもはや単なる UI ではなく、制約を強制し、行動を監視し、人間の介入を可能にする環境でなければならない。AI が「数日間にわたって自律的に行動する」[1] ためには、アプリケーション層が堅牢なインフラストラクチャを提供する必要がある。これには、ツールへの安全な API アクセス、異常行動の監視、説明責任のためのロギング、および人間の監督または「キルスイッチ」のメカニズムが含まれる [30]。LlamaFirewall [16] および LLM 向け OWASP トップ 10 [16] の議論は、アプリケーション層が緩和を支援しなければならないセキュリティ脆弱性を強調している。

OpenAI が「エージェンシーの前の推論」[6] を強調していることは、信頼性の高い自律的行動には強力な認知的基盤が必要であることを意味する重要な戦略的選択である。これは、改善された推論なしに現在のエージェント的アプローチを単にスケーリングするだけでは、信頼性の低いまたは安全でないシステムにつながるという信念を示唆している。[6] および [6] の記述は、OpenAI がレベル 3（エージェント）の前にレベル 2（推論者）を優先することを明確に述べている。これは、「Devin」[6] のようなエージェントを既存のモデルを使用して構築しようとする一部のスタートアップとは対照的である。OpenAI のスタンスは、エージェントが長期間にわたって確実に計画、適応、自己修正する能力は、その中核的な推論能力に根本的に依存することを示唆している。これがなければ、エージェントはより頻繁なエラーや「関数呼び出しハルシネーション」[37] を起こしやすくなる可能性がある。

D. レベル 4：イノベーター（創造的 AI）

1. 能力と特性の定義

レベル 4 の AI システムは、新しい発明、アイデア、創造的な解決策を自律的に生成できると定義される [4]。問題解決を超えて、独創的な発想と発見へと移行する [6]。研究開発、科学的発見（例：病気の治療）、創造的プロジェクト（例：芸術、音楽）への応用

が期待される[1]。このレベルのAIは、関連タスクにおいて熟練した人間の99%を上回るパフォーマンスを発揮するとされる[1]。創造性と独創的思考において人間の能力を反映し、AGIへの大きな飛躍を表す[7]。

2. LLM層の役割と進化

クリティカルなLLM機能

- **抽象的概念化と類推：** 抽象的な概念を形成・操作し、異なるドメイン間で類推を行う能力。
- **創造的仮説生成：** 新規で検証可能な仮説を策定する能力[27]。
- **分野横断的知識統合：** 多様な分野からの情報を統合し、新たな洞察を生み出す能力[18]。
- **制約のない探求：** 人間のセレンディピティに似て、事前に定義された制約なしに解決策空間を探求する能力[18]。
- **新規性と有用性の評価：** 生成されたアイデアの独創性と潜在的価値を評価する能力。

LLMアーキテクチャ/トレーニング

より探索的で制約の少ない生成プロセスをサポートするアーキテクチャが必要となる可能性がある。トレーニングには、科学文献、特許、芸術作品の膨大なデータセット、および実験のためのシミュレートされた環境への暴露が含まれる場合がある。新規性追求型の報酬を用いた強化学習や進化的アルゴリズムなどの技術が役割を果たす可能性がある。仮説が科学的に妥当であることを保証するための知識ベースの思考連鎖（KG-CoI）アプローチ[39]。

3. アプリケーション層の役割と進化

アプリケーションタイプ

AI駆動のR&Dプラットフォーム、自動科学発見ツール、創造的コンテンツ生成スイート（例：芸術、音楽、文学向け）、創薬プラットフォーム[27]。

LLMの活用

アプリケーション層は、研究問題や創造的目標を定義するためのインターフェースを提供し、革新的なLLMとの対話を管理し、AIが生成したアイデアのテスト、検証、改良を促進する。実験室の自動化、シミュレーションツール、または大規模なデータセットと統合される場合がある。

複雑性

非常に複雑であり、堅牢なデータ管理、実験計画能力、新規概念のための高度な視覚化ツール、および潜在的に人間と AI の共同イノベーションのための協調環境が必要となる。

4. 主要な技術的ハードルと新たな倫理的考慮事項

技術的ハードル

- **「新規性」と「創造性」の定義と評価：** 真のイノベーションを定量化することは困難である [18]。
- **自律的な問題再定義：** AI は現在、問題空間を自律的に再定義したり、仮定に異議を唱えたりするのに苦労している [18]。
- **根拠付けと実現可能性：** AI が生成したイノベーションが科学的に有効であるか、物理的に実現可能であるか、または実用的に有用であることを保証すること。
- **誘導と制御：** 創造性を阻害することなく、AI を有用なイノベーションに向けて誘導すること [18]。
- **イノベーションのためのマルチモーダル統合：** より豊かな創造的統合のために、多様なデータタイプをシームレスに統合すること [18]。

倫理的考慮事項

- **知的財産：** AI が生成した発明や創造的作品の所有権 [42]。
- **イノベーションにおけるバイアス：** AI が新しい発見や技術にバイアスを永続させたり導入したりする可能性がある [31]。
- **予期せぬ結果：** AI が生成した新規のアイデアは、予測不可能で潜在的に否定的な社会的影響を与える可能性がある [38]。
- **人間のイノベーターの役割：** 人間の科学者、芸術家、発明家への影響 [27]。
- **AI 生成科学の信頼性：** 厳密性と再現性を保証すること（[50]、[18]、[18]、[18]、[40]、[50] – ただし [50] と [18] はアクセス不可だが、その質問はこれらの懸念を示している）。
- **有害なイノベーションに対する説明責任：** [31]。

「イノベーター」は、AI が人間定義のタスクを実行するためのツールから、創造的および発見プロセスのパートナー、あるいは扇動者へと根本的に移行することを表す。これは、R&D と創造性そのものの性質に深遠な影響を与える。レベル 1〜3 は、人間がすでに行う方法を理解しているタスク（会話、推論、行動）を AI が実行することに焦点を当てているが、潜在的により効率的である。レベル 4 の「イノベーター」は、AI が「新しい発明やアイデア」[4] を生成することを意味する。これは単なるタスク実行ではなく、知識創造である。これにより、人間と AI の関係は、ユーザーとツールから、共同研究者と共同研究者へ、あるいは AI が真に新しい洞察を推進する場合、指導者と弟

子へと変化する。

「イノベーター」向けアプリケーション層は、高度な「AI R&D コパイロット」または「発見プラットフォーム」へと進化する必要がある。これらのプラットフォームは、AI の出力だけでなく、真のイノベーションに必要な実験的検証、データ統合、人間と AI の共同ワークフローも管理する必要がある。アイデアの生成（LLM 機能）はイノベーションの最初のステップにすぎない。アイデアはテストされ、改良され、文脈化される必要がある。アプリケーション層（例：創薬向け[27]）は、シミュレーションツール、実験用ハードウェア（ラボオートメーション）、および既存の知識のデータベースと統合して、このライフサイクルをサポートする必要があるだろう。また、人間の専門家が AI の革新的な出力を指導、検証、解釈するためのインターフェースも必要となる。

レベル 4 における倫理的課題、特に IP および偏ったまたは有害なイノベーションの可能性に関するものは、AI が新規性を生成しているため、以前のレベルよりも著しく複雑である。既存の法的および倫理的枠組みは不十分である可能性がある。AI が新しい薬や材料を発明した場合[27]、誰が特許を所有するのか[42]？AI が偏ったデータに基づいて新しい社会政策を提案した場合、現実世界の害が発生する前にそのバイアスはどのように特定され緩和されるのか[31]？現在の IP 法は人間中心である。AI がこれらのイノベーションをどのように導き出すかという「ブラックボックス」の性質[31]は、説明責任と信頼性を非常に困難にする[18]。

E. レベル 5：組織（AGI/ASI プロキシ）

1. 能力と特性の定義

このフレームワークにおける AI 開発の頂点であり、AI システムが組織全体の集合的な機能を実行できる段階である[4]。複雑なワークフローを管理し、複数のプロジェクトを大規模に監督し、戦略的思考、運用効率を処理し、組織目標を達成するために適応する[1]。このレベルの AI は、関連タスクにおいて人間の 100%を上回るパフォーマンスを発揮するとされる[1]。AI がほとんどの経済的に価値のあるタスクで人間の能力を凌駕する AGI の実現としばしば同一視される[5]。システムは完全に自律的であり、動的に学習し自己改善する[23]。

2. LLM 層の役割と進化

クリティカルな LLM 機能

- **包括的システム理解：** 複雑で相互接続されたシステムと組織のダイナミクスの深い理解。
- **戦略的推論と長期計画：** 組織規模での複数年にわたる戦略の策定と実行、リソー

ス配分、リスク管理[27]。
- **複数 AI エージェント/サブシステムの調整：** 集合的な目標を達成するために、多様な AI コンポーネントを調整する能力[27]。
- **大規模な倫理的意思決定：** 組織の文脈において複雑な倫理的枠組みを組み込み適用する能力[27]。
- **継続的な自己改善と適応：** 組織のパフォーマンスと環境の変化から学習し、運用を自律的に最適化する能力[23]。

LLM アーキテクチャ/トレーニング

非常に複雑で、潜在的にハイブリッドな AI アーキテクチャが関与する可能性が高く、専門化された LLM や他の AI モジュールのネットワークで構成される場合がある。トレーニングには、ビジネスオペレーション、経済学、社会ダイナミクス、および潜在的にシミュレートされた組織環境を網羅する膨大で多様なデータセットが必要となるだろう。継続的な学習と適応メカニズムがアーキテクチャの中核となるだろう。

3. アプリケーション層の役割と進化

アプリケーションタイプ

完全に自律的な AI 経営の企業または組織単位[4]。AI CEO、AI マネージャー、AI 駆動の運用プラットフォーム[4]。

LLM の活用

アプリケーション層は AI 組織そのものである。それは、中核となる AI インテリジェンス（LLM 層）がすべての組織機能を管理および実行する構造、プロセス、およびインターフェースを具体化する。これには、残りの人間の監督または高レベルの目標設定のためのインターフェースが含まれる。

複雑性

AI アプリケーションアーキテクチャにおける究極の複雑性を表し、戦略計画から顧客との対話、R&D、リソース管理に至るまで、組織の運用のあらゆる側面を、まとまりのある AI 駆動システムに統合する[33]。

4. 主要な技術的ハードルと新たな倫理的考慮事項

技術的ハードル

- **極度の複雑性管理：** 組織全体の複雑さを持つ AI システムを調整および制御すること。

- **堅牢性と回復力：** AI 組織が予期せぬ混乱や障害に耐えられることを保証すること。
- **大規模な人間価値との整合性（スーパーアラインメント）：** AI 組織の目標と行動が人間の意図と社会的利益と整合し続けることを保証することが最も重要である[45]。
- **解釈可能性とデバッグ：** このような広大で複雑な AI システム内の問題を理解し修正すること。

倫理的考慮事項

- **実存的リスク：** 誤った方向に調整された超知能 AI が壊滅的な危害を引き起こす可能性[36]。
- **権力集中：** AI 組織は、経済的および意思決定権の前例のない集中につながる可能性がある[42]。
- **社会的混乱：** 大量失業、経済構造の根本的な変化[4]。
- **ガバナンスと制御：** このような強力な AI エンティティはどのように統治され制御されるのか？誰が責任を負うのか？[45]。
- **公平性と公正性：** AI 組織の利益が公平に分配され、不平等を悪化させないことを保証すること。
- **セキュリティ：** AI 組織を高度な攻撃や内部の誤動作から保護することは、世界的なセキュリティの問題となる[20]。プロンプトインジェクションやデータポイズニングなどのリスクは、AI が運営する組織を危険にさらす可能性がある場合、実存的な脅威となる。

レベル 5 の「組織」は、AGI が個々の認知タスクだけでなく、人間社会の制度に似た、複雑で多面的で目標指向のシステムを長期にわたって調整・管理する能力に関するものであることを示唆している。これは、例えば一連の認知テストに合格するだけよりもはるかに高いハードルである。「組織全体の集合的な機能を実行する」[4] および「複雑なワークフローを管理し、複数のプロジェクトを大規模に監督する」[4] という記述は、個々の知能を超える能力を示している。これには、通常、人間の組織におけるリーダーシップと管理に関連するシステム思考、戦略計画、リソース配分、適応などの機能が含まれる。これは、OpenAI の見解では、AGI がこれらのメタ認知機能と実行機能を含むことを示唆している。

レベル 5 における「アプリケーション層」は、本質的に組織自体、つまり AI ネイティブなエンティティとなる。AI（LLM/中核的知能）とアプリケーション（組織構造とプロセス）の区別は著しく曖昧になり、完全に統合される可能性がある。AI が「複雑なワークフローを管理」し、「複数のプロジェクトを監督」[4] している場合、「アプリケ

ーション」はもはや AI が使用する個別のソフトウェアではなく、運用のまさに枠組みである。AI は CEO であり、マネージャーであり、運用システムである[4]。これは、AI の推論が AI 定義および AI 実行プロセスを通じて組織の行動に直接変換される、深く統合されたシステムを意味する。

この段階では、「スーパーアラインメントイニシアチブ」[45] が絶対に不可欠となる。AI が組織を運営できる場合、その目標は人間の価値観と証明可能かつ堅牢に整合していなければならない。なぜなら、誤った方向に調整された AI「組織」からの大規模な負の影響の可能性は計り知れないからである。AI 組織[4] は大きな力を持つだろう。その中核となる LLM によって駆動される内部目標が、人間の利益からわずかでも逸脱した場合、有害な方法でそれらの目標を最適化する可能性がある（例：極端な社会的コストで利益を最大化する）。スケーラブルな監視と堅牢性テストに焦点を当てたスーパーアラインメントイニシアチブ[45] は、AI がこのレベルの能力と自律性に達したときに最も深刻になるこの実存的リスクを直接認識したものである。

IV. AGI に向けた LLM 層とアプリケーション層の共進化と深化する統合

A. 高度化と相互依存の強化

LLM の能力が基本的な NLP（レベル 1）から複雑な推論（レベル 2）、自律的な計画とツール使用（レベル 3）、創造的な革新（レベル 4）、そして包括的な組織管理（レベル 5）へと進化する過程を概観する。同時に、アプリケーション層が単純な UI（レベル 1）から問題解決のための高度なプラットフォーム（レベル 2）、エージェントオーケストレーション（レベル 3、[14]）、R&D とイノベーション管理（レベル 4）、そして最終的には AI 駆動の組織そのもの（レベル 5）へと進化する軌跡をたどる。一方の層における進歩が、他方の層における進歩を直接的に可能にし、必要とすることを強調する。例えば、より強力な LLM 推論（レベル 2）は、より複雑な問題に取り組むことができるアプリケーションを可能にする。自律エージェント（レベル 3）の必要性は、計画とツール使用における LLM 開発を推進する。

B. 並行的な進歩の必要性

AGI は、LLM 開発またはアプリケーションエンジニアリングのいずれかに単独で焦点を当てることでは達成できないと論じる。真の AGI は、両者のシームレスで高度に統合された機能から生まれる。LLM は「心」を提供するが、アプリケーション層は「身体」と「感覚」（ツール統合、データフィードを通じて）および「手」（行動実行能力を通じて）を提供する。

C. 統合のためのアーキテクチャパラダイム

複雑さが増すにつれて、モノリシックな AI システムからよりモジュール化され分離されたアーキテクチャへの移行について議論する。特にレベル 3〜5 において、LLM アプリケーション向けの提案された 3 層分離アーキテクチャ（アプリケーション層、プロトコル層、ハードウェア層）[14] の関連性を強調する。

- **アプリケーション層：** アプリケーション構成、マルチエージェントオーケストレーション、クロスプラットフォーム相互運用性を処理する。
- **プロトコル層：** 効率的なタスク実行、安全な通信、標準化されたデータ交換を保証する。
- **ハードウェア層：** 特殊な AI アクセラレータを含む、基礎となる計算を提供する。このようなアーキテクチャが、非常に複雑な AGI レベルのシステムを構築および管理するために不可欠なモジュール性、再利用性、スケーラビリティ、相互運用性、および保守性をどのように向上させることができるかを説明する。

3 層アーキテクチャ [14] における「プロトコル層」は、AI がレベル 3（エージェント）以降へと進むにつれてますます重要になる。それは、高度な LLM が複雑なアプリケーション環境内で多数のツールや他の AI エージェントと確実に相互作用することを可能にする「神経系」である。レベル 1 および 2 の AI は、より単純なアプリケーション-LLM 接続で動作するかもしれない。しかし、レベル 3 のエージェントは、外部ツールや潜在的に他のエージェントとの堅牢な相互作用を必要とする [4]。レベル 4 のイノベーターは、実験を制御したり、多様なデータセットにアクセスしたりする必要があるかもしれない。レベル 5 の組織は、相互作用するコンポーネントの広大なネットワークを伴うだろう。標準化され、効率的で、安全なプロトコル層は、この複雑な通信とオーケストレーションに不可欠であり、アドホックな統合の混沌とした増殖を防ぐ。

共進化はフィードバックループを意味する。LLM がより複雑なアプリケーションを可能にするにつれて、これらのアプリケーションは、今度はさらに有能な LLM をトレーニングするために使用できる新しいタイプのデータと相互作用パターンを生成する。これは、AI 開発を加速させる好循環（または、誤って調整された場合は潜在的に危険な）を生み出す。レベル 3（エージェント）の LLM は、その相互作用とタスクの成功/失敗から学習する（適応の必要性から推測される）。これらの相互作用からのデータ（例：成功したツール使用シーケンス、エージェントのパフォーマンスに関するユーザーフィードバック）は、LLM をより優れたエージェントにするためのさらなるトレーニングに非常に価値がある。同様に、イノベーターAI（レベル 4）は、アプリケーション層ツールを介してテストされる新しい仮説を生成する可能性があり、これらのテストの結果は LLM の知識にフィードバックされる。この反復的な改良が進歩の鍵となる。

LLM 層とアプリケーション層がより深く統合され、有能になるにつれて、セキュリティ脆弱性や倫理的侵害の「攻撃対象領域」も拡大し、より複雑になる。リスクはもはや LLM のみ、またはアプリケーションのみに限定されず、それらの相互作用から生じる。プロンプトインジェクション [16] は、LLM とアプリケーションのインターフェースにおけるリスクの典型的な例である。アプリケーション層における安全でないプラグイン設計 [21] は、LLM を操作にさらす可能性がある。エージェント（レベル 3）がより多くの自律性とツールアクセスを得るにつれて、LLM の意思決定またはアプリケーションのツールアクセスコントロールのいずれかの脆弱性が、「過剰な権限付与」[16] と重大な危害につながる可能性がある。これには、システム全体を考慮した包括的なセキュリティアプローチが必要となる。

V. 広範な課題と将来展望

A. 包括的な技術的課題

- **スケーラビリティ：** 計算リソース、トレーニングと推論のためのデータ要件。
- **堅牢性と信頼性：** 特に新しい状況において、AI システムが期待どおりに一貫して動作することを保証すること。
- **説明可能性と解釈可能性：** 非常に複雑な AI の意思決定プロセスを理解すること。
- **継続的な学習と適応：** 壊滅的な忘却や不安定性を伴わずに、AI が現実世界の環境で学習し適応できるようにすること。
- **イノベーションのための現在の生成 AI の限界：** 問題を自律的に再定義する能力の欠如、真の制約のないイノベーションメカニズムの欠如 [18]。

B. 深刻化する倫理的および社会的課題

- **アラインメント問題：** AGI の目標が人間の価値観と意図と一致することを保証すること [17]。これは多くの倫理的議論の中心的なテーマである [25]。
- **バイアスと公平性：** AI が社会のバイアスを大規模に永続させたり増幅させたりするのを防ぐこと [17]。
- **プライバシーとデータセキュリティ：** 高度な AI によって処理される膨大な量のデータを保護すること [22]。
- **誤用と悪意のある行為者：** 高度な AI が兵器化されたり、有害な目的で使用されたりするのを防ぐこと [22]。
- **雇用喪失と経済的影響：** 広範な AI 自動化によって引き起こされる社会的変化に対処すること [4]。
- **説明責任と責任：** 高度な AI システムが危害を引き起こした場合に、明確な説明責任の所在を確立すること [17]。

C. ガバナンス、安全性、協力の重要性

堅牢なガバナンスフレームワークと規制の必要性[17]。IEEEの原則（人権、責任、透明性、誤用の緩和）が出発点となる[17]。安全性研究、敵対的テスト、制御メカニズム開発への重点[45]。AGI への道筋をナビゲートする上での国民参加と国際協力の役割[45]。OpenAI の憲章は、別のプロジェクトが AGI に近づいている場合、競争よりも協力を優先することに言及している[6]。

「アラインメント問題」は単一の課題ではなく、AI の能力が増大するにつれて指数関数的に困難になる一連の問題である。チャットボット（レベル 1）を協力的で無害なものに調整することは、AI 組織（レベル 5）を動的で制約のない環境における多様な人間の価値観と整合させることとは、規模と複雑さの点で異なる。[17]および[17]は価値観の埋め込みについて議論している。レベル 1 では、これは有害なコンテンツのフィルタリングを意味するかもしれない。レベル 5 では、「倫理的意思決定能力」[27]および「スーパーアラインメント」[45]は、AI の中核業務への複雑な倫理的推論のより深く体系的な統合を意味する。「道徳的過負荷」のリスク（価値観の対立、[17]）は、組織を管理する AI にとって計り知れないものとなる。

堅牢なグローバルガバナンスと安全プロトコルによって抑制されない AGI への競争は、競争圧力が個々の行為者に安全性を軽視させ、集団的リスクにつながる「コモンズの悲劇」シナリオにつながる可能性がある。[46]および[46]は、地政学的競争と倫理的考慮事項が脇に追いやられるリスクを強調している。ある主体が、優位性を得るために高度に有能だが不十分に調整された AI（例：レベル 4 または 5 のシステム）を展開した場合、不安定な状況を生み出し、他者に安全性が保証される前に自らの展開を加速させることを強いる可能性がある。これは、[45]および[46]で言及されている国際協力と強制力のある基準の決定的な必要性を強調している。

「AGI」の定義とそれへの認識されている近接性（サム・アルトマンのタイムライン対他の専門家[1]）は、投資、研究の方向性、および国民の認識に大きな影響を与える。主要なプレイヤーが差し迫ったブレークスルーを予測する場合、より保守的で安全性を重視したアプローチが損なわれ、より迅速に動くよう圧力がかかる可能性がある。サム・アルトマンの楽観的なタイムライン[1]は、より慎重な専門家の見解[1]と対照的である。この見通しの違いは、安全対策がどれほど緊急に実施されるか、対して能力がどれほど迅速に推進されるかに影響を与える可能性がある。「競争に勝つ」[46]ことに焦点が当てられている場合、各レベルでのアラインメントを確保し倫理的課題に対処するという複雑な作業は、リソース不足になるか、急がされる可能性がある。

VI. 結論

A. 役割の要約

OpenAI の AGI5 段階レベル全体にわたる、LLM 層（言語処理から包括的知能へ）とアプリケーション層（単純なインターフェースから複雑で自律的な運用環境へ）の、明確でありながら絡み合った進化の道筋を簡潔に要約する。

B. AGI に向けた絡み合った道筋

AGI への進歩は、単により強力な LLM やより高度なアプリケーションに関するものだけではなく、それらの相乗的な共進化と深化する統合に関するものであることを改めて強調する。

C. 総括

AGI への道のりに関連する計り知れない可能性と深遠な課題について、将来を見据えた声明で締めくくり、責任ある開発、堅牢な倫理的枠組み、および積極的なガバナンスの決定的な必要性を強調する。

引用文献

1. OpenAI's 5 phases toward AGI | Next Level Impact, 5 月 18, 2025 にアクセス、https://nxtli.com/en/openai -agi/
2. curriculumredesign.org, 5 月 18, 2025 にアクセス、https://curriculumredesign.org/wp -content/uploads/The -Frustrating-Quest-to-Define-AGI.pdf
3. What Is Artificial General Intelligence and Why It Matters to Business | PYMNTS.com, 5月 18, 2025 にアクセス、https://www.pymnts.com/artificial -intelligence-2/2025/what-is-artificial-general-intelligence-and-why-it-matters-to-business/
4. The path to AGI: A deep dive into openAI's 5-level framework - Twimbit, 5 月 18, 2025 にアクセス、https://twimbit.com/about/blogs/the -path-to-agi-a-deep-dive-into-openais-5-level-framework
5. What Are OpenAI's Five Levels of AI-- And Where Are We Now?, 5 月 18, 2025 にアクセス、https://theaiinsider.tech/2024/07/12/what -are-openais-five-levels-of-ai-and-where-are-we-now/
6. Understanding OpenAI's Five Levels of AI Progress Towards AGI, 5月 18, 2025 にアクセス、https://quinteft.com/understanding -openais-five-levels-of-ai-progress-towards-agi/
7. OpenAI's 5 Steps to AGI: Level 2 Reasoners Officially Here- First Movers, 5 月 18, 2025 にアクセス、https://firstmovers.ai/openai -agi-levels/
8. What is a large language model (LLM)? | SAP, 5月 18, 2025 にアクセス、https://www.sap.com/resources/what -is-large-language-model

9. What is a large language model (LLM)? - SAP, 5 月 18, 2025 にアクセス、 https://www.sap.com/swiss/resources/what-is-large-language-model
10. What Are Large Language Models (LLMs)? - IBM, 5 月 18, 2025 にアクセス、 https://www.ibm.com/think/topics/large-language-models
11. Guide to Large Language Models (LLMs) Explained | Balbix, 5 月 18, 2025 にアクセス、 https://www.balbix.com/insights/what-is-large-language-model-llm/
12. What is the Application Layer? A Deep Dive into Network Communication - OurCrowd, 5 月 18, 2025 にアクセス、 https://www.ourcrowd.com/learn/what-is-the-application-layer
13. What is the Application Layer? - AI Glossary, 5 月 18, 2025 にアクセス、 https://www.theainavigator.com/blog/what-is-the-application-layer
14. The Next Frontier of LLM Applications: Open Ecosystems and Hardware Synergy - arXiv, 5 月 18, 2025 にアクセス、 https://arxiv.org/pdf/2503.04596?
15. arxiv.org, 5 月 18, 2025 にアクセス、 https://arxiv.org/pdf/2503.04596
16. Taming the Machine: Putting Security at the Core of Generative AI - Tidal Cyber, 5 月 18, 2025 にアクセス、 https://www.tidalcyber.com/blog/taming-the-machine-putting-security-at-the-core-of-generative-ai
17. standards.ieee.org, 5 月 18, 2025 にアクセス、 https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf
18. arxiv.org, 5 月 18, 2025 にアクセス、 https://arxiv.org/html/2503.11419v1
19. From Generative AI to Innovative AI: An Evolutionary Roadmap - arXiv, 5 月 18, 2025 にアクセス、 https://arxiv.org/pdf/2503.11419?
20. AI Risks: Exploring the Critical Challenges of Artificial Intelligence | Lakera – Protecting AI teams that disrupt the world., 5 月 18, 2025 にアクセス、 https://www.lakera.ai/blog/risks-of-ai
21. The Complete Guide to AI: Application Security Testing - Checkmarx, 5 月 18, 2025 にアクセス、 https://checkmarx.com/learn/appsec/the-complete-guide-to-ai-application-security-testing/
22. What are the ethical concerns surrounding OpenAI? - Milvus, 5 月 18, 2025 にアクセス、 https://milvus.io/ai-quick-reference/what-are-the-ethical-concerns-surrounding-openai
23. The Path to AGI: How Do We Know When We're There? - Lumenova AI, 5 月 18, 2025 にアクセス、 https://www.lumenova.ai/blog/artificial-general-intelligence-measuring-agi/
24. Understanding OpenAI's Five Levels of AI Progress Towards AGI ..., 5 月 18, 2025 にアクセス、 https://quinteft.com/understanding-openais-five-levels-of-ai-progress-towards-agi
25. Ethics of AI - Put your ethical concerns here - Page 2 - Community ..., 5 月 18, 2025 にアクセス、 https://community.openai.com/t/ethics-of-ai-put-your-ethical-concerns-here/1119333?page=2
26. ChatGPT 101: The Student Ethics Edition - by Liza Long - Artisanal Intelligence, 5 月 18, 2025 にアクセス、 https://lizalong.substack.com/p/chatgpt-101-the-

student-ethics-edition?utm_source=post-email-title&publication_id=2411318&post_id=163739613&utm_campaign=email-post-title&isFreemail=true&r=h74tq&triedRedirect=true&utm_medium=email
27. AI Breakthroughs: Emotions & Ethics | We Create GPTs, 5 月 18, 2025 にアクセス、 https://gpt.gekko.de/openai-strawberry-reached-level-2/
28. [The AI Show Episode 145]: OpenAI Releases o3 and o4-mini, AI Is ..., 5 月 18, 2025 にアクセス、 https://www.marketingaiinstitute.com/blog/the-ai-show-episode-145
29. Introducing OpenAI o3 and o4-mini | OpenAI, 5 月 18, 2025 にアクセス、 https://openai.com/index/introducing-o3-and-o4-mini/
30. 5 Levels of agentic AI intelligence for enterprise use - Outshift, 5 月 18, 2025 にアクセス、 https://outshift.cisco.com/blog/agentic-ai-intelligence-for-enterprise-use
31. Balancing Innovation and Responsibility: The Challenge of Ethical AI, 5 月 18, 2025 にアクセス、 https://www.nextwealth.com/blog/balancing-innovation-and-responsibility-the-challenge-of-ethical-ai/
32. Ethics of AI - Put your ethical concerns here - Page 5 - Community ..., 5 月 18, 2025 にアクセス、 https://community.openai.com/t/ethics-of-ai-put-your-ethical-concerns-here/1119333?page=5
33. cdn.openai.com, 5 月 18, 2025 にアクセス、 https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf
34. Professional Agents - Evolving Large Language Models into Autonomous Experts with Human-Level Competencies - arXiv, 5 月 18, 2025 にアクセス、 https://arxiv.org/html/2402.03628v1
35. The Five Stages of AI Agent Evolution - NFX, 5 月 18, 2025 にアクセス、 https://www.nfx.com/post/ai-agent-revolution
36. arxiv.org, 5 月 18, 2025 にアクセス、 https://arxiv.org/pdf/2503.04750
37. New Ethics Risks Courtesy of AI Agents? Researchers Are on the ..., 5 月 18, 2025 にアクセス、 https://www.ibm.com/think/insights/ai-agent-ethics
38. OpenAI's Quest for Artificial General Intelligence: Vision or Mirage?, 5 月 18, 2025 にアクセス、 https://casmi.northwestern.edu/news/articles/2024/openais-quest-for-artificial-general-intelligence-vision-or-mirage.html
39. Embracing Foundation Models for Advancing Scientific Discovery ..., 5 月 18, 2025 にアクセス、 https://pmc.ncbi.nlm.nih.gov/articles/PMC11923747/
40. 1 月 1, 1970 にアクセス、 https://www.lakera.ai/blog/risks-of-ai/
41. GPT-4.1: OpenAI unveils its latest model - Swiftask, 5 月 18, 2025 にアクセス、 https://www.swiftask.ai/blog/gpt-4-1
42. Exploring the Ethical Implications of OpenAI: Balancing Innovation ..., 5 月 18, 2025 にアクセス、 https://3cloudsolutions.com/resources/exploring-ethical-implications-open-ai/
43. OpenAI's Five-Step Journey from AI to AGI - HulkApps, 5 月 18, 2025 にアクセス、 https://www.hulkapps.com/blogs/ecommerce-hub/openais-five-step-

journey-from-ai-to-agi
44. 7 reasons why AI reasoning models accelerate time to value ..., 5 月 18, 2025 にアクセス、 https://lumenalta.com/insights/7-reasons-why-ai-reasoning-models-accelerate-time-to-value
45. General - Artificial General Intelligence | NMU AI Literacy Initiative, 5 月 18, 2025 にアクセス、 https://nmu.edu/ai-literacy-initiative/general-intelligence-ai
46. DeepSeek and the Race to AGI: How Global AI Competition Puts ..., 5 月 18, 2025 にアクセス、 https://www.techpolicy.press/deepseek-and-the-race-to-agi-how-global-ai-competition-puts-ethical-accountability-at-risk/
47. OpenAI's 5 Steps to AGI - Perplexity, 5 月 18, 2025 にアクセス、 https://www.perplexity.ai/page/openai-s-5-steps-to-agi-STzklF5SSQ6JOiBTaV.cfA
48. OpenAI outlines plan for AGI—5 steps to reach superintelligence | Tom's Guide, 5 月 18, 2025 にアクセス、 https://www.tomsguide.com/ai/chatgpt/openai-has-5-steps-to-agi-and-were-only-a-third-of-the-way-there
49. Retrieval Augmented Generation in Production with Haystack, 5 月 18, 2025 にアクセス、 https://4561480.fs1.hubspotusercontent-na1.net/hubfs/4561480/Ebooks%20whitepapers%20and%20reports/O%E2%80%99Reilly%20Guide%20-%20RAG%20in%20Production%20with%20Haystack/OReilly%20Guide%20-%20RAG_in_production_with_Haystack-FINAL.pdf
50. An Evaluation of Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards 'Artificial General Research Intelligence' (AGRI)? - arXiv, 5 月 18, 2025 にアクセス、 https://arxiv.org/html/2502.14297v1