

Thinking Beyond Tokens: From Brain-Inspired Intelligence to Cognitive Foundations for Artificial General Intelligence and its Societal Impact

Kazuyuki Yoshikawa

<https://github.com/anas-zafar/agi-cognitive-foundations>

要旨 Abstract

- 機械（AI）は本当に人間のように考え、推論し、行動できるのか？これは人工汎用知能（AGI）の追求における永続的な問いである。GPT-4.5、DeepSeek、Claude 3.5、Sonnet、Phi-4、Grok 3のようなモデルは、マルチモーダルな流暢性や部分的な推論能力を示しているが、それらは基本的にトークンレベルの予測（token-level prediction）に依存しており、grounded agency（基盤のある主体性／行動主体性）が欠けている。
- 本論文ではAI、認知神経科学、心理学、生成モデル、エージェントベースシステムといった学際的観点からAGI開発を総合的に分析する。
- モジュール型推論、永続的メモリ、多エージェント協調の役割を強調する。
- Agentic RAG（Retrieval, Planning, Tool Use を組み合わせ、より適応的な行動を可能にする枠組み）の台頭にも注目。
- 情報圧縮、テスト時適応、トレーニングフリーな手法など汎用性のある知性に向けた一般化戦略を議論する。
- VLM（Vision-Language Models）は単なる知覚モジュールとしてではなく、具現化（embodied）された理解および共同タスク遂行のためのインターフェースとして再検討される。
- 真の知性はスケール（モデルサイズなど）のみからは生まれない。メモリと推論を統合した、モジュラーで対話的、自律改善が可能な構成が必要であり、圧縮が適応的振る舞いを可能にする。
- 最後に、AGIへの道における科学的・技術的・倫理的課題を特定する

背景 Background

- 現在の大規模言語モデル（LLMs）やマルチモーダルモデルは、驚異的な成果を挙げているが、「トークン予測（次の語を予測する）」ことに大きく依存しており、人間のような「目的を持った行動」「持続的メモリ」「推論や計画」「主体性 (agency)」の面で限界があると主張されている。
- 認知科学・神経科学の成果から、人間の知性・知覚・学習・思考にはモジュール性（複数の異なる機能モジュールの存在）、記憶の永続性、環境との相互作用、多様な情報の統合、社会的・協調的な要素などが重要。これらを AGI の開発に応用する必要がある。

方法 Methods／アプローチ

- 横断的統合 (cross-disciplinary synthesis) :
- AI、認知神経科学、心理学、生成モデル、エージェントベースシステムの知見を整理
- 構造的／アーキテクチャ的要素の分析 :
- モジュラー推論、永続メモリ、マルチエージェント協調などがどうAGIに寄与するかを分析。
- 汎化戦略の考察：情報圧縮 (information compression)、テスト時適応 (test-time adaptation)、トレーニングフリー手法 (training-free methods) などを取り上げ、モデルが未知ドメインでも適応可能であるための戦略を議論
- 具体的なモデル枠組みの紹介：Agentic RAG フレームワーク (Retrieval, Planning, Tool Use を組み合わせたもの) など。

主な結果 Results／提案

- 現行モデルの限界点を整理し、スケールだけでは**AGI**には不十分であるとの結論。特に、「記憶と推論の統合」が鍵であるとしている
- **Agentic RAG** のような枠組みが、より柔軟で目的指向の行動をモデルに持たせる上で有力であることを示唆。
- **Vision-Language Models** を単なる入力-出力型の知覚モジュールとして捉えるのではなく、環境との動的なやりとりを含む「具現化された理解 (embodied understanding)」を担うものとして設計する意義を論じている
- また、「情報圧縮」「テスト時適応」など、未知条件下で性能を維持・向上させるための戦略が重要であるという指摘。

議論 Discussion

- AGI に向けては、ただモデルを巨大化するだけでなく、知性を構成する複数の要素（記憶、推論、計画、主体性、環境との相互作用など）を統合することが必要
- また、汎用性・適応力という観点から、未知の環境やドメインでの性能を保つためには、訓練時以外でも適応できる構造／戦略が求められると主張

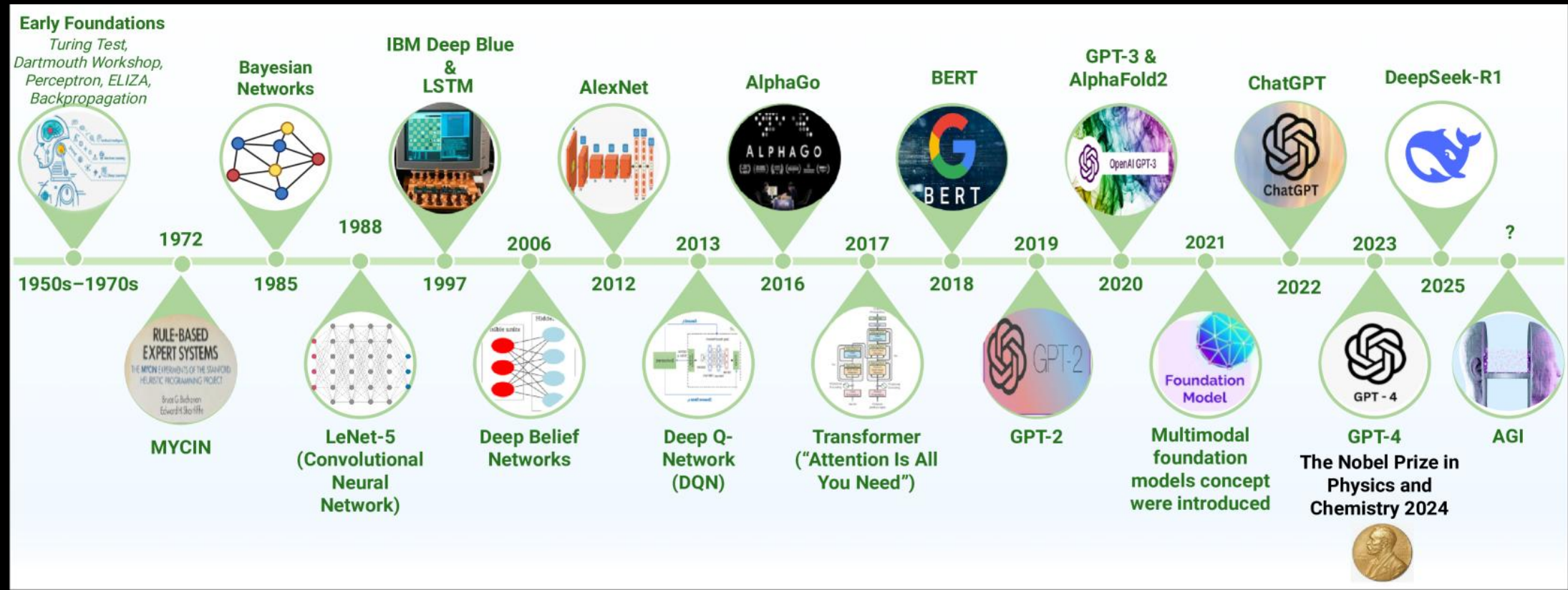
この研究の新規性 Novelty

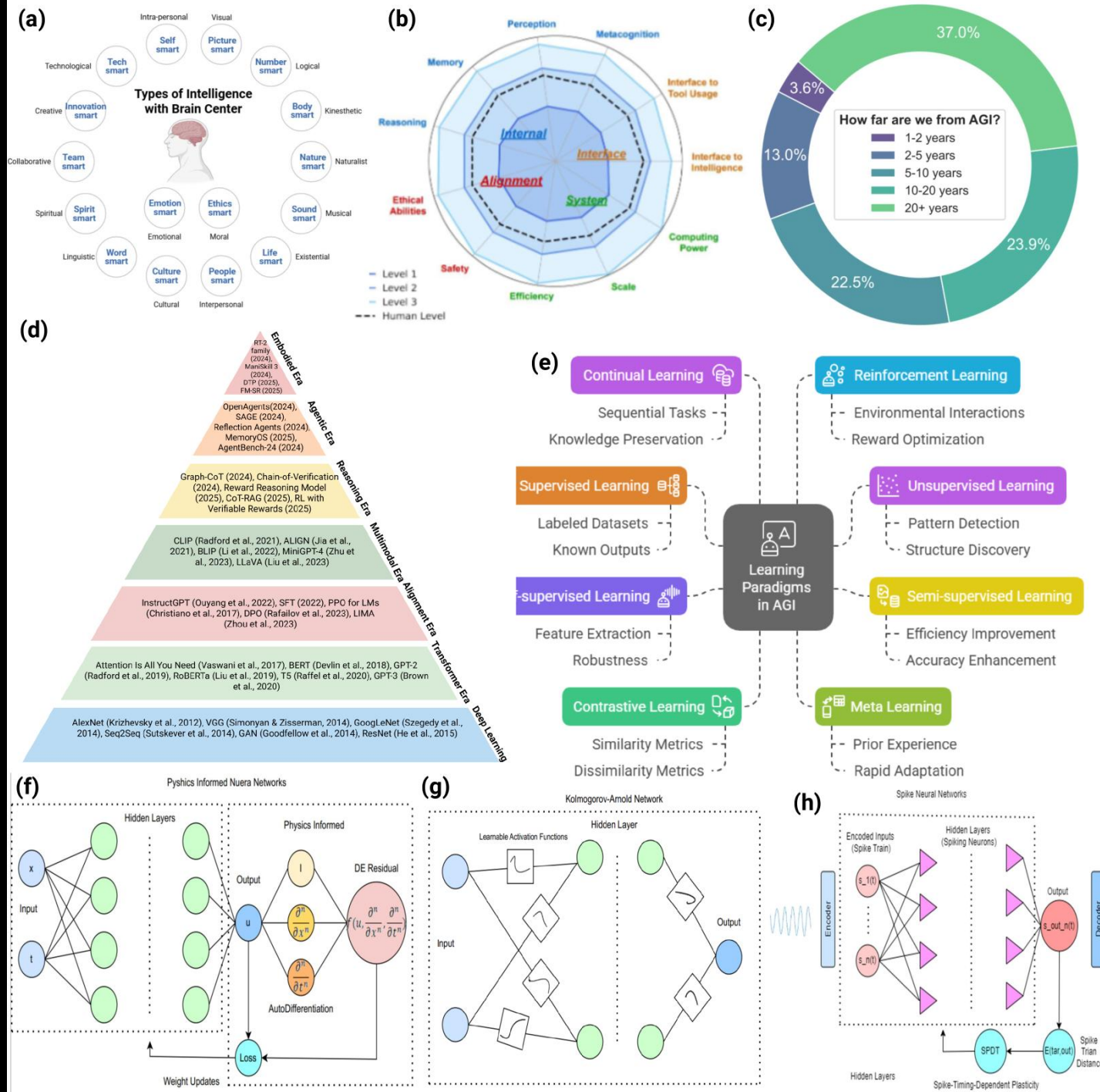
- 「トークン予測 (token - level prediction)」 中心のモデルからの脱却を明確に標榜し、より認知科学・神経科学的な知見を取り込んだ AGI アーキテクチャの基盤を整理した点。
- **Agentic RAG** のような枠組みを取り上げ、それを **AGI** の実現に向けて重要な要素と位置づけたこと。
- 訓練フェーズだけでなく、推論時（またはその後）での適応や汎化能力に焦点を当てている点。
- **Vision-Language Models** を具現化 (embodied) の視点から再評価するところなど、単なる入力-出力モデルとしてではない知性の形を議論している点。

定義

- **思考**:問題を解決し、世界について推論し、新しいアイデアを生み出すための内部表象の操作[2].
- **意識**:認識と内省のための主観的な能力[4].
- **インテリジェンス**:タスクや環境全体で知識を習得、適用、適応させる能力[3].
- **AGI**:タスク固有の再トレーニングを必要とせずに、ドメインを超えて、人間レベルの幅広い推論と学習が可能なシステム[5].

- 図 1:1950 年から 2025 年までの汎用人工知能 (AGI) に向けた主要なマイルストーンのタイムライン。
- シンボリックシステム(例:ELIZA)、ニューラルネットワーク(例:LeNet-5、AlexNet)、強化学習(例:AlphaGo、DQN)、基盤モデル(例:GPT-4、DeepSeek-R1)に及びます。この軌道は、静的なルールベースの手法から、動的でマルチモーダルでますます一般的な AI システムへの移行を反映しています。



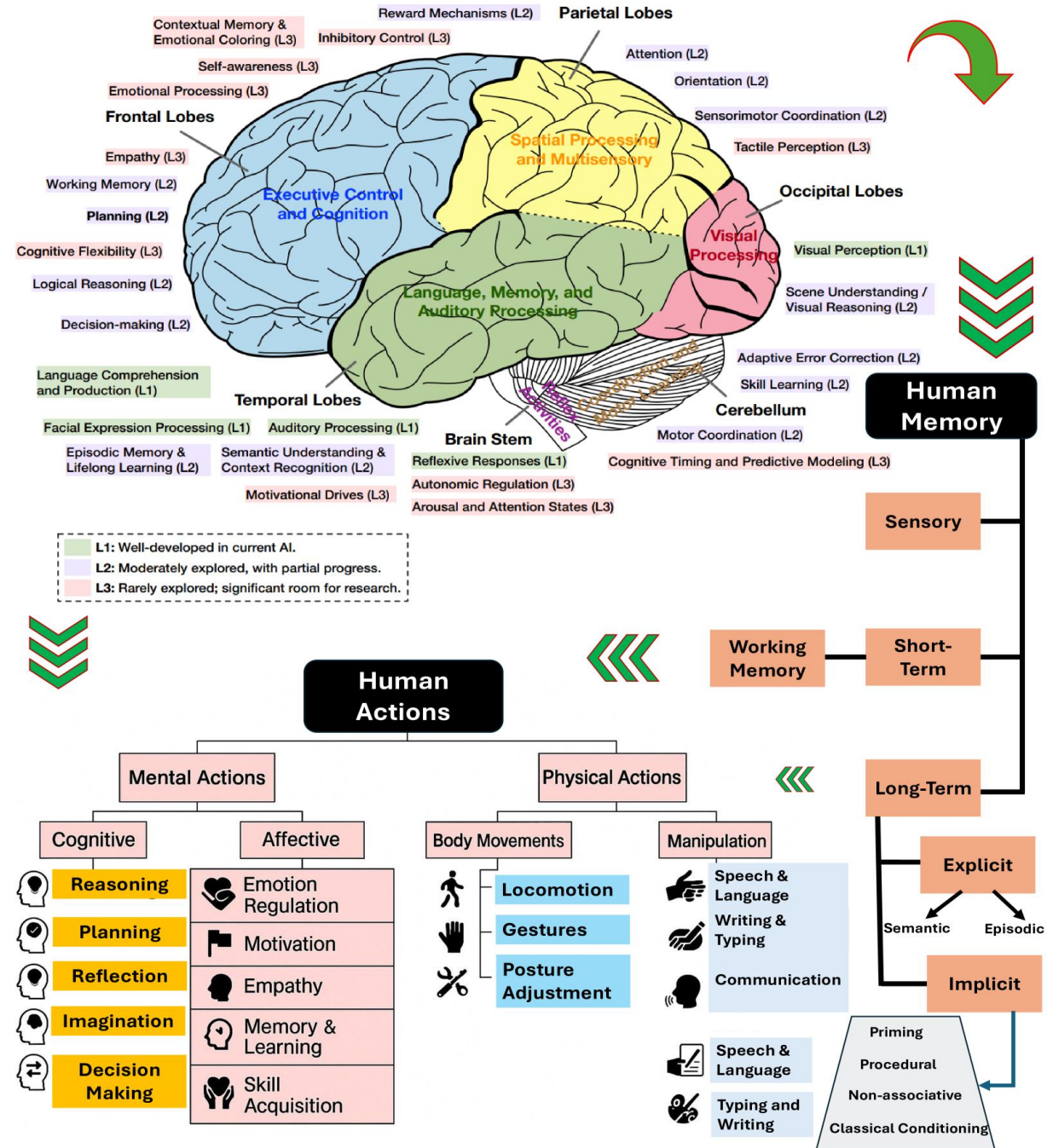


- 図 2:汎用人工知能 (AGI) に向けた基本的な概念、進歩、パラダイムの概要。
- (a) 脳に触発されたAGIで概念化された複数の人間の知能タイプ。
- (b) 内部推論、外部インターフェイス、システム効率、倫理的安全性など、AGIにおける多次元アライメントの課題を表すレーダーチャート。
- (c) ICLR 2024調査から採用したAGIタイムライン期待の調査ベースの予測[\[26\]](#).
- (d) 具現化された時代につながる基礎的な AI 時代のピラミッド。
- (e) 教師あり学習、教師なし学習、自己教師あり学習、強化学習、および継続学習、対照学習、半教師あり学習、メタ学習などの新たなパラダイムを含む、AGIにおけるコア学習パラダイムの分類。
- (f) 物理情報に基づくニューラルネットワーク (PINN)、
- (g) コルモゴロフ-アーノルドネットワーク (KAN)
- (h) スパイクニューラルネットワーク (SNN) を表すアーキテクチャは、AGI 開発における生物学的妥当性と適応計算を強調しています。

脳の機能

- 哺乳類の進化の特徴である新皮質は、高次の認知と抽象的な推論
- 皮質下構造は感情機能と自律神経機能を調節
- 海馬などの主要なコンポーネントはエピソード記憶 (EM) と空間ナビゲーションのエンコードを促進
- 後頭皮質は視覚処理を制御し、
- 運動皮質は随意運動を調整
- 約 150 兆のシナプス接続からなる密なネットワークを作成する
約 860 億個のニューロン

- 前頭葉は、計画や意思決定などの高レベルの認知を司る[95]、AIは構造化されたタスク (AlphaGo など) で強力なパフォーマンスを示します。
- 意識や認知の柔軟性などの特性は、まだ十分に研究されていません (L3)
- 対照的に、L1 ドメインにマッピングされた言語および聴覚機能は、言語処理における人間レベルの習熟度に近づく LLM によって適切にモデル化されています
- 逆に、小脳系と大脳辺縁系は、それぞれ細かい運動能力と感情処理を支配します
- AI では、ロボット工学とメタ学習を通じて運動協調が探求されます。しかし、人間のような器用さと適応性を実現することは依然として課題です (L2-L3).
- 大脳辺縁系によってモデル化された感情的および動機付けのプロセスは、強化学習を通じて AI で表面的にしか再現されておらず、真の心の知能指数の開発における大きなギャップが浮き彫りになっています。 (L3)




- 図 4: 正確なマルチステップの視覚的推論を実行する際の主要な大規模マルチモーダルモデル(LMM)の限界の図解。構造化された思考の連鎖に従うように促されたにもかかわらず、すべてのモデルは、既知の部屋の高さ 257 cm に基づいてクリスマスツリーの高さを正しく推定できません。

Input


Context: Think step by step and use chain-of-thought reasoning to answer the following question.

Question: What is the height of the Christmas tree knowing that the height of the room (floor to ceiling) is 257 cm? Give your answer in cm to the nearest 5cm.

Input Image



Ground Truth: '190 cm'

**Gemini-2.5 Pro**

Answer: 230 cm

Reasoning: A visual estimation of the Christmas tree's height in relation to the room's height is needed to answer this. Here's the thinking process:

Step 1: Analyze the image: Observe the Christmas tree and its proximity to the ceiling.


Step 2: Estimate the tree's proportion to the room height.

Step 3: Calculate the estimated height range:

- Lower end: $257 \text{ cm} \times 0.85 = 218.45 \text{ cm}$
- Upper end: $257 \text{ cm} \times 0.95 = 244.15 \text{ cm}$

Step 4: A 90% estimate of the 257 cm room height gives 231.3 cm. Rounding to the nearest 5 cm gives 230 cm.

Step 5: This fits the visual gap of about 27 cm from the ceiling, making 230 cm a reasonable estimate.

**InternVL-3-14B**

Answer: 165 cm

Reasoning:


Step 1: Room Height: The total room height from floor to ceiling is 257 cm.

Step 2: Tree Proportion: Christmas trees are generally not as tall as the room. They might reach about 1 to 2/3 of the room height, depending on the room size and tree size.

Step 3: Tree in Image: The tree in the image has a decorative object at its base (a hat) and is standing in a corner. The tree appears to be more than half the room height but not reaching the ceiling.

$$x = \frac{2}{3} \times 257 \approx 171.3 \text{ cm}$$

Step 4: Visual Verification: Given the positioning and proportions in the image, the tree appears slightly shorter than this estimation. Thus, the height is about 165 cm.

**Qwen2.5-VL-32B**


Answer: 206-231 cm

Reasoning: To determine the height of the Christmas tree, we need to analyze the image and use proportional reasoning based on the given information about the room's height.

Step 1: Observe the Image: The tree stands against the wall, shorter than the ceiling, with small figures nearby for scale.

Step 2: Estimate Proportion: The tree appears to cover about 80%–90% of the room's height.

Step 3: Calculate Height: With a room height of 257 cm, the tree is roughly 206–231 cm tall.

**GPT-4o**

Answer: 195 cm

Reasoning: To determine the height of the Christmas tree, we follow these steps:

Step 1: The real height of the room is given as 257 cm.

Step 2: The image height in pixels is 865 px.

Step 3: The Christmas tree spans approximately from pixel 110 to pixel 770, giving it a pixel height of 660 px.

Step 4: Using proportional scaling, we calculate the tree's real-world height:

$$x = \frac{660}{865} \times 257 \approx 195 \text{ cm}$$

Step 5: Rounding to the nearest 5 cm gives us:

Final Answer: The height of the Christmas tree is approximately 195 cm.

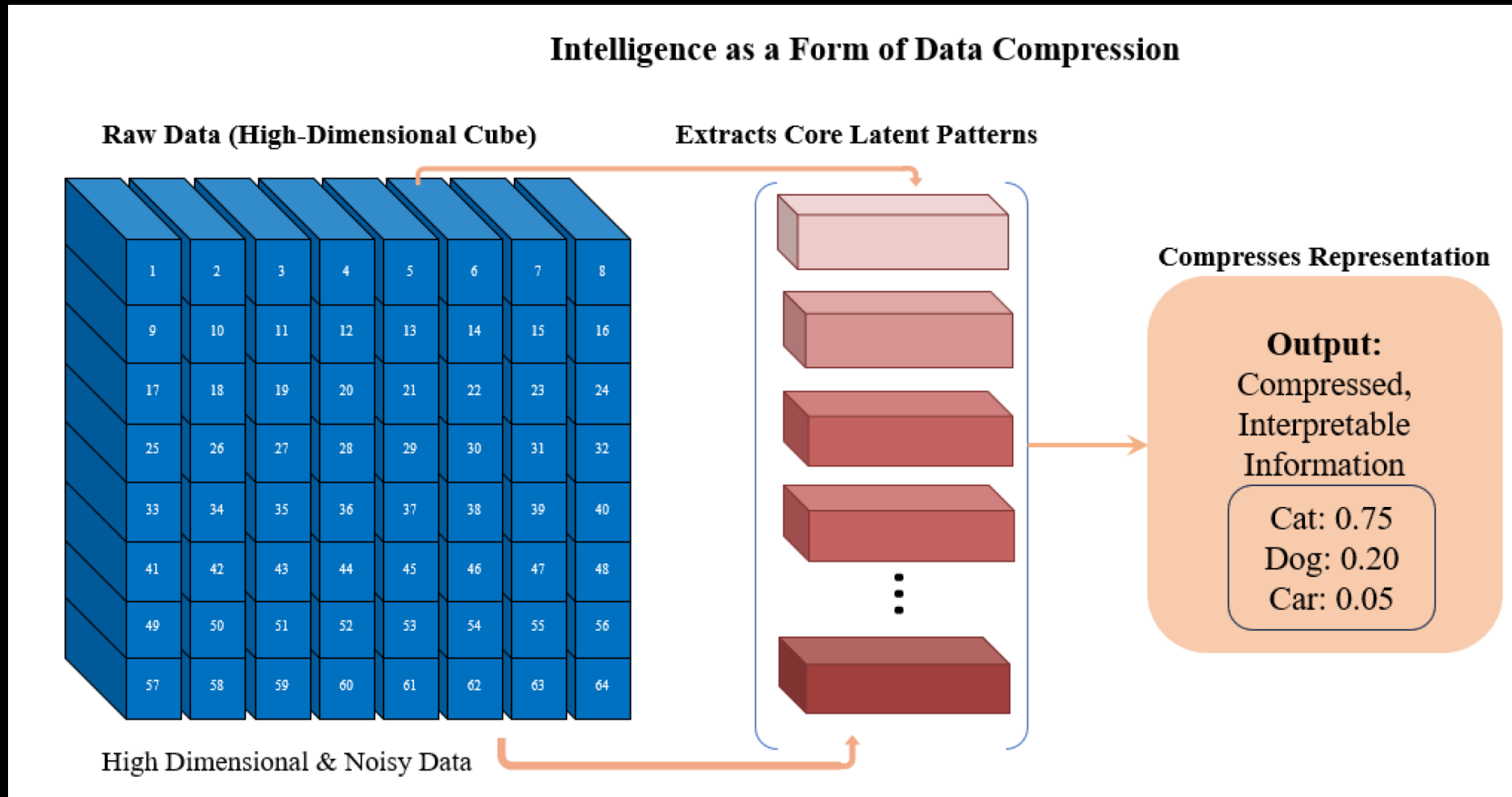
世界モデル

- ワールドモデルは、エージェントが試行錯誤だけに依存することなく、シミュレーション、予測、計画を行えるようにする内部表現です
- 人間では、これらのメンタルモデルは空間ナビゲーション、計画、反事実推論を支えます、予測的、適応的、スケーラブルな認知を提供します[125].
- 予測は、視覚的な手がかりと事前の経験を統合し、知覚と記憶によって洗練されます。アクションはAIのような意思決定モジュールを通じて選択され、フィードバックはメモリと内部モデルを更新します。
- この図は、4つの概念層にまたがって構成されています:
- (1) 基本的な世界モデルタイプ(暗黙的、明示的、シミュレーターベース、命令駆動型)。
- (2) 予測、階層、およびフィードバックによる動的推論。
- (3) 中核となる主体能力の知覚、記憶、および行動。
- (4) 倫理的推論や文脈適応性を含む意欲的な AGI 能力。

表1:AGI研究における脳機能に触発されたニューラルネットワーク

後頭葉	ビジュアル処理	畳み込みニューラルネットワーク(CNN)	画像認識、物体検出	生物学的視覚は、まばらな階層的なフィルタリングを使用します。CNN は、エッジとテクスチャにレイヤー フィルターを適用します
海馬/側頭葉	メモリエンコーディング、シーケンスモデリング	リカレントニューラルネットワーク (RNN) 、LSTM	シーケンシャルモデリング、時系列予測	人間は適応的に文脈を思い出します。RNNは限られた時間状態を捉える
運動皮質	自発的なモーションコントロール	ロボット制御ネットワーク	ロボット工学、運動能力学習	人間の動きは固有受容感覚とフィードバックを使用します。ロボットのポリシーは最適化に依存しています
前頭前野	計画と意思決定	強化学習(RL)	ゲームプレイ、ナビゲーション、戦略タスク	人間は不確実性と価値観の下で計画を立てます。RLは報酬の最大化に焦点を当てています
シナプス可塑性	時間的ダイナミクスによる学習	スパイクニューラルネットワーク(SNN)	ニューロモーフィックモデリング、リアルタイム推論	ヘブ/STDP ルールは人間の学習を導きます。SNN は、スケーラビリティのトレードオフを伴うスパイクをシミュレートします
聴覚皮質	言語と音声の理解	トランスフォーマーネットワーク	言語モデリング、翻訳、テキスト生成	人間は感情と文脈を統合します。トランスフォーマーはシーケンスに対してトークンアテンションを使用します

- 図 5: 圧縮としてのインテリジェンスの例: ノイズの多い入力(左)は、潜在的な抽象化(中央)と単純化された出力(右)に抽出され、一般化と推論が強化されます。



パネルA:神経に着想され、物理学に基づいたアーキテクチャ

アーキテクチャ	SNN	PINN	従来のNN
財産	スパイクタイミングとイベント駆動型シグナリングのシミュレーション	ニューラルユニット内の物理的制約のエンコード	トレーニング可能な重みを使用した抽象的な人工ニューロン
タイムダイナミクス	スパイクによる時間的エンコーディング	タスク駆動型の暗黙的な時間表現	RNNを使用しない限り、しばしば存在しない
計算パラダイム	イベントベースのエネルギー効率の高い処理	偏微分方程式制約付きデータ近似	データ駆動型汎用マッピング
生物学的アライメント	高(可塑性、スパース性)	中程度(物理リアリズム)	低(柔軟だが抽象的)
効率	適度;最適化	ソルバーの複雑さに依存	高スループット/GPU並列処理
ユースケース	エッジロボティクス、動的センシング	科学シミュレーション、気候モデリング	視覚、NLP、強化学習
AGIの可能性	リアルタイムの認識	物理学によるシンボルのグラウンディング	スケーラブルなパターン抽象化

パネルB:一般化のための理論的構成

理論	誘導原理	財団	AGIへの影響
情報のボトルネック (IB)	ノイズを破棄しながら、関連する潜在機能に焦点を当てる	情報論、相互情報	コンパクトでタスク関連の表現学習
最小記述長 (MDL)	シンプルさは一般化を支持します	アルゴリズム情報理論	圧縮された解釈可能なモデルを選択します
暗黙的正則化 (SGD)	最適化中のフラット最小値	損失ランドスケープ ジオメトリ	一般化を奨励する
NTK / ダブルディセント	過剰パラメータ化されたレジームは、後期の一般化に利益をもたらします	無限幅カーネル理論	堅牢な学習の体制を特徴付ける
PAC-ベイズバウンド	分布事前確率からの一般化	確率的学習理論	形式一般化の保証
因果表現学習	介入に不変の安定した因果関係の特徴を抽出	因果グラフ、SEM	タスク/ディストリビューション全体の堅牢性を促進
変分ドロップアウト	学習したノイズインジェクションによる正則化	変分推論	スパース性とノイズ耐性を強化
シンプルさバイアス	より単純な仮説を最初に学習します	トレーニングの経験的ダイナミクス	トレーニングの初期段階での複雑さの軽減

パネルC:学習アルゴリズムと損失関数バイアス

機構	誘導バイアス	例	AGIとの関連性
SGD / 早期停止	より平坦な最小値に対する暗黙的な優先	クラシクトレーニングのセットアップ	一般化可能で安定した収束
アダプティブ・オプティマイザー (Adam、RMSProp)	収vergenceは速いが、鋭いソリューションのリスクがある	LLM の微調整、低データ設定	速度と一般化のトレードオフ
クロスエントロピー損失	自信のある予測を促進	分類タスク	シンプルでありながら不確実性に鈍感
コントラスト/トリプレット損失	潜在クラスタリング、関係構造	SimCLR、MoCo、トリプレットネット	堅牢な表現学習
KLダイバージェンス(VAE、PACベイズ)	潜在空間または分布を正則化します	VIB、ベイズネットワーク	最小限のもつれのないコードを奨励します
RLの目的	長期単位の割り当て、目標の焦点	PPO、Qラーニング、DPO	計画と逐次推論をサポート
メタラーニング / PEFT	タスクに依存しない初期化または高速適応	MAML、LoRA、爬虫類	効率的な少数ショットまたは継続的な学習が可能

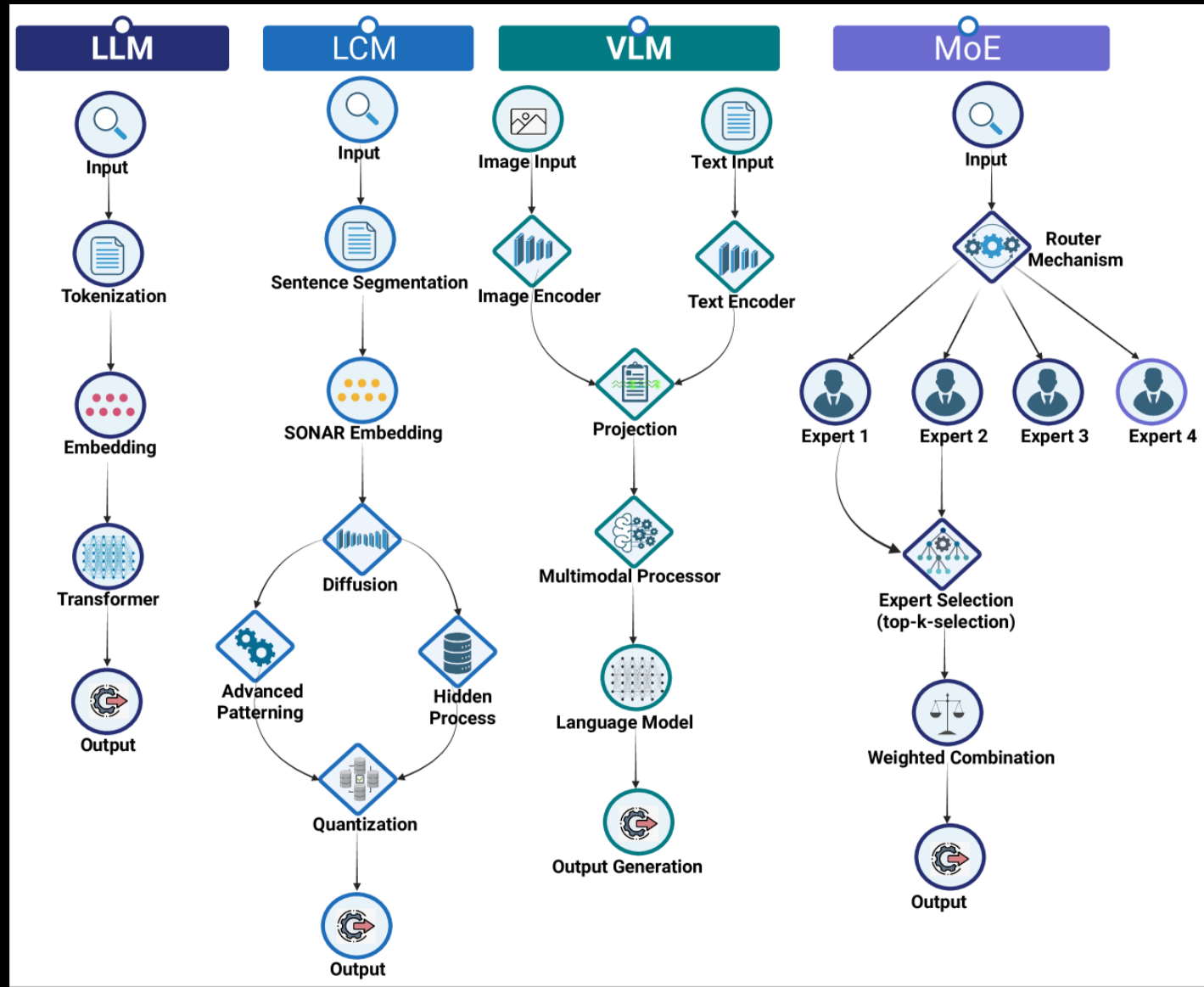
パネルD:基盤モデルにおける新たな帰納的事前確率

機構	誘導バイアス	例	AGI の関連性
マルチモーダルアテンション	モダリティ間の整合性が可能	CLIP、フラミンゴ、Perceiver IO	根拠のある推論と知覚的理解をサポートします
クロスモーダル対照学習	共有構造を介して視覚的および言語の埋め込みを調整	アライメント、LiT、GIT	共有表現と構成性を奨励します
外部メモリの拡張	長期的かつ一時的な想起を促進します	RNN+メモリ、リアクト、レトロ	スケーラブルなコンテキストとシンボリック・チェーンが可能
検索拡張生成 (RAG)	推論中の外部データベース	RAG、アトラス、KAT	事実性/適応性を強化
マスクモデリング/自己帰納	部分的なコンテキストから予測構造を学習します	BERT、GPT、BEiT、MAE	汎用自己教師あり事前トレーニング
プロンプトチューニングと命令バイアス	タスクのプロンプトや指示を通じて構造を学習します	T5、InstructGPT、PEFT、プレフィックスチューニング	ゼロショットの適応とユーザーの意図への調整を提供します
ヒューマンフィードバックによるRL(RLHF)	モデル出力を人間の価値観/好みに合わせる	InstructGPT、DPO、憲法AI	安全性と価値の整合性に不可欠

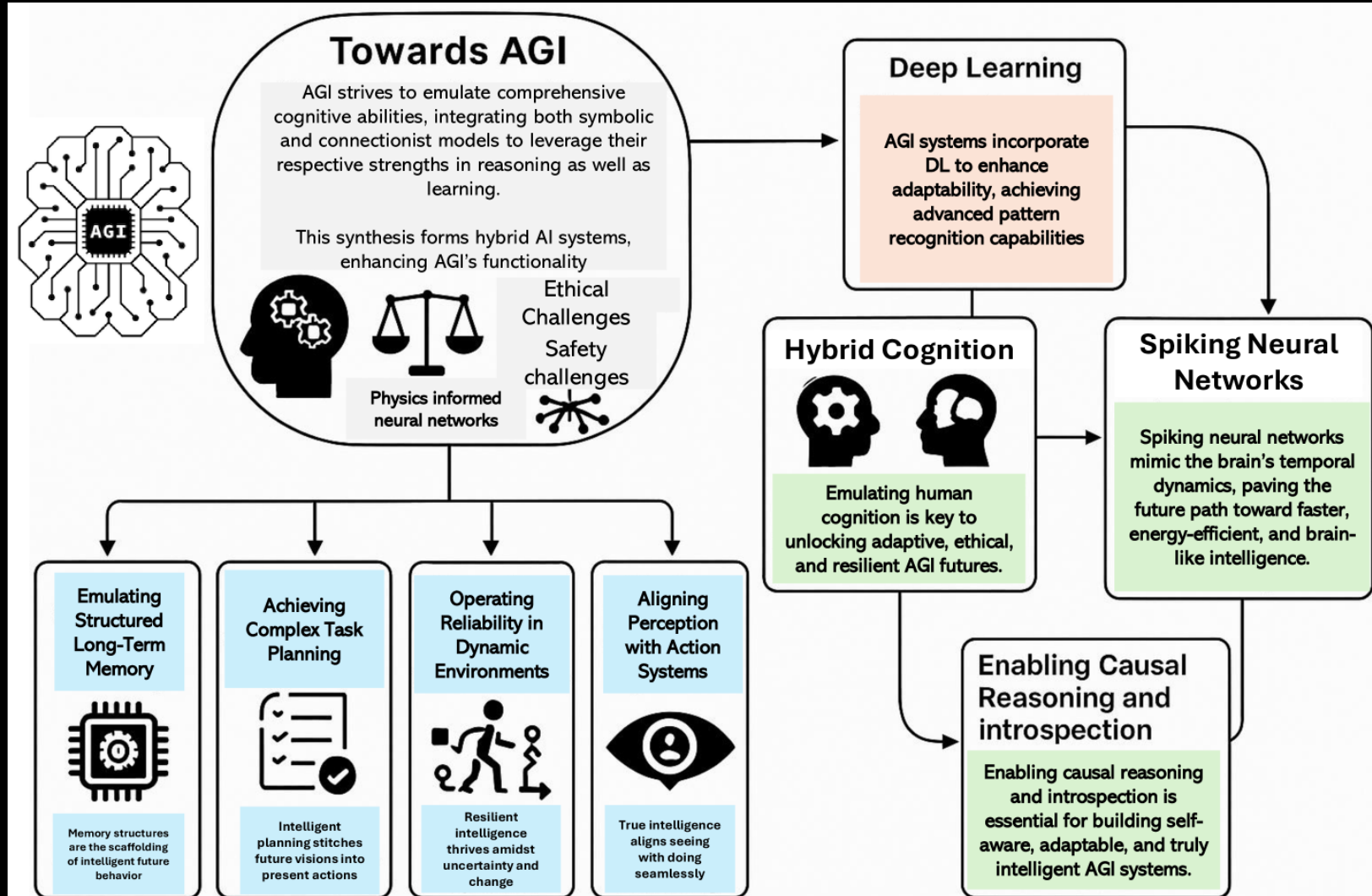
コア認知機能

- 推論
 - AGI システムは、新しい問題を解決するために演繹的、帰納的、および帰納的推論を実行する必要があります[240,35].深い推論により、仮説の検証、計画、反事実推論が可能になります[241].思考連鎖や神経記号システムなどのモデルは、記号論理と神経学習を統合して、より解釈可能で適応的な推論を実現します[242,243,244].
- 7.1.2学習
 - AGI は、教師あり、教師なし、シンボリック、強化、深層学習のパラダイムを統合します[245,246]. これらにより、一般化と継続的な改良が可能になります。強化学習は、動的な環境での相互作用ベースの学習を促進します[247]、ディープラーニングはモダリティ全体で特徴を抽象化します[248].
- 7.1.3思索
 - 思考とは、抽象化、戦略形成、意思決定を指します。認知アーキテクチャとニューラルネットワークは、高レベルの思考をシミュレートします[249]. 神経記号システムは、形式論理と適応可能なモデルを組み合わせたものです[250]、複雑な推論タスクの信頼性の向上[251].
- 7.1.4記憶
 - 記憶は、コンテキスト認識と学習の継続性をサポートします。短期記憶は、即時のタスク処理に役立ちます。長期記憶は蓄積された知識をコード化します[78,252]. パラメトリックおよび外部メモリシステムにより、迅速な検索と柔軟な更新が可能[71].
- 7.1.5知覚
 - AGI 知覚には、マルチモーダルな感覚解釈が含まれます。CNNとトランスフォーマーは、視覚信号と聴覚信号を処理します[253].Perceiver や Flamingo などのマルチモーダル モデルの進歩により、異種入力を解釈する AGI の能力が向上します[254].

- 図 7: コア基盤モデルアーキテクチャの概念的概要。大規模言語モデル (LLM)、言語中心モデル (LCM)、ビジョン言語モデル (VLM)、および Mixture of Experts (MoE) のアーキテクチャパイプライン。



- 図 8: シンボリックモデルとコネクショニストモデルを統合し、構造化記憶、因果推論、適応計画、知覚と行動の調整を強調し、将来の開発のための安全性、効率性、内省的な認知能力に取り組むことで、人間のような知能へのAGIの進歩を示しています。



AGI実現への鍵：脳からのインスピレーション

- モジュール性: 脳には視覚を処理する「後頭葉」、記憶を司る「海馬」など、機能ごとに特化した領域があります。
- 同様に、AIも単一の巨大なモデルではなく、特定の機能を持つ**専門家のAI（モジュール）**を組み合わせる「専門家混合（Mixture of Experts, MoE）」アーキテクチャが有効だと述べています。
- 記憶システム: 人間の記憶は、短期記憶から長期記憶へと情報を整理・統合します。
- AIにも、単なる情報保持だけでなく、文脈に応じて情報を統合し、柔軟に活用できる高度な記憶システムが必要だと指摘しています。
- 神経科学との連携: 脳活動を測定するfMRIなどの技術を用いて、人間が思考や意思決定を行う際の脳のネットワーク活動を分析し、それをAIの設計に応用することの重要性を強調しています。

エージェント型AI (Agentic AI)

- 従来のAIが「指示に応答する」受動的な存在だったのに対し、
- エージェント型AIは自ら目標を設定し、計画を立て、ツールを使いこなし、他のエージェントと協調してタスクを遂行します。
- これは、静的な知識を持つ「予測マシン」から、動的な世界で行動する「知的エージェント」へのパラダイムシフトを意味します。

トークンを超えるモデル

- 大規模概念モデル (LCMs): 単語単位ではなく、文全体の「概念（コンセプト）」レベルで情報を処理することで、より人間らしい深い文脈理解を目指すモデル
- 大規模推論モデル (LRMs): 答えを即座に出すのではなく、人間のように「じっくり考える」プロセスをモデル化し、多段階の推論を経て結論を導き出します。
- ニューロシンボリックAI: ニューラルネットワークのパターン認識能力と、記号論理学の厳密な推論能力を組み合わせ、双方の長所を活かすハイブリッドなアプローチです

「般化」の重要性

- 真の知能は、学習したことを単に記憶するのではなく、その知識を応用して未知の新しい問題を解決する能力（般化）にあります。
- 論文では、「知能とは、本質を捉えて情報を効率的に圧縮する能力である」という情報ボトルネック理論などを紹介し、表面的なパターン学習からの脱却を促しています
- アライメントと安全性: **AGI**が人間の価値観や意図から外れた行動を取らないように制御する「アライメント」が不可欠です。そのために、人間のフィードバックを取り入れながらAIを学習させる**RLHF** (Reinforcement Learning with Human Feedback) などの技術が重要になります。