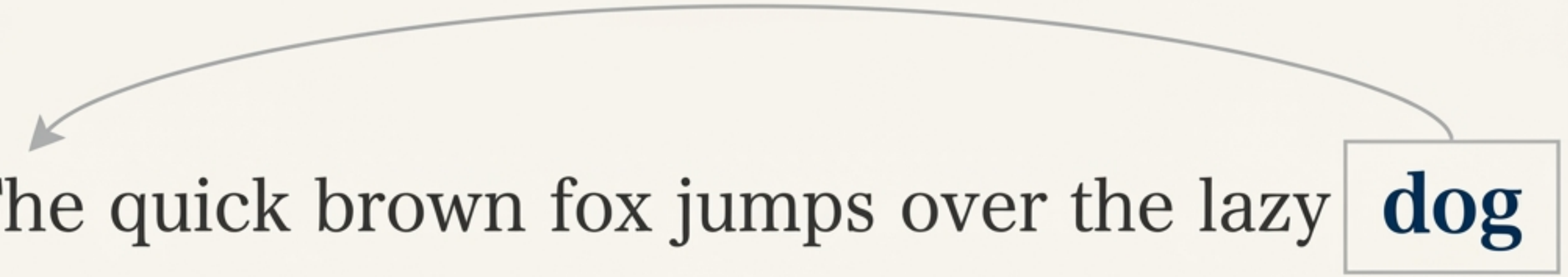
LLMの次章：単なる言語モデルから「真のワールドモデル」へ

テキスト予測の限界を超え、世界を理解するAIへのロードマップ

テキスト予測の限界を超え、
世界を理解するAIへのロードマップ

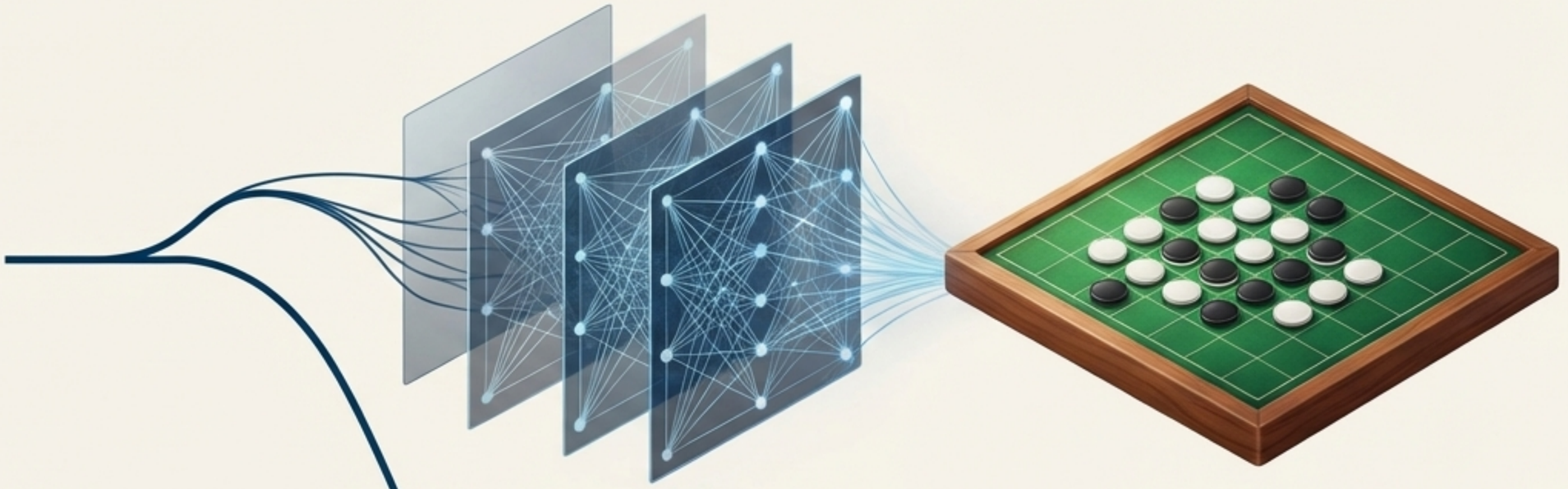
「大規模言語モデルは、単語を予測するだけなのに、なぜ世界を「理解」しているように見えるのか？」



技術的には、LLMは自己回帰型の次トークン予測という単純なメカニズムで動作する。
しかし、その内部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分かってきた。
このスライドでは、その「創発的能力」の根源と、それが示唆する「内部ワールドモデル」の存在を探る。

内部世界の発見：Othello-GPTが示した創発的な盤面表現

f5,
d6,
c5,
e6,
f6...



タスク：次の合法手を予測

言語モデルにオセロの棋譜のみを学習させ、「次の合法手」を予測するよう訓練。モデルには盤面のルールを一切教えていない。

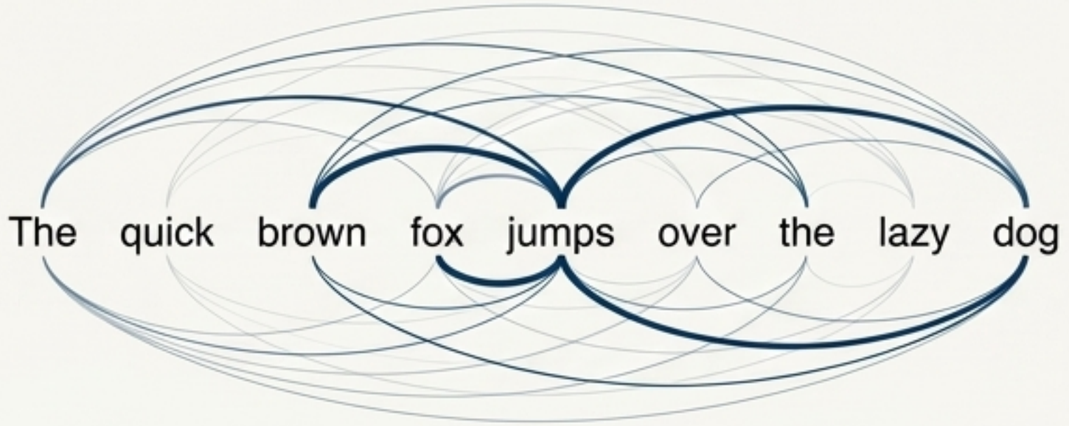
結果： 内部に盤面状態のベクトル表現が自発的に形成

驚くべきことに、モデルの中間層のベクトルから、現在の石の配置置（盤面状態）を線形変換で正確に読み出すことができた。これは、モデルが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、「ゲームの世界モデル」を獲得したことを示唆する。

世界の断片はどのように符号化されているか

自己注意機構（Self-Attention）

トランスフォーマーの核となる機能。文脈中の関連する単語間の関係性を動的に重み付けし、長距離の依存関係を捉えることで、一貫した世界シミュレーションを可能にする。低次の語句から高次の抽象概念へと、階層的に表現を洗練させる。



知識ニューロン（Knowledge Neurons）

モデル内部には、特定の事実関係に対応するニューロンが存在することが確認されている。「アゼルバイジャンの首都はバクー」といった事実に特定のニューロンが発火し、これを操作するとモデルの回答を改変できる。これはLLMが事実知識を分散表現として内包している証拠である。

「アゼルバイジャンの首都は？」

「人間レベルの知能に至る道としては行き止まりである」

Yann LeCun（Meta チーフAIサイエンティスト）

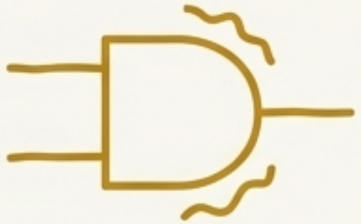
テキストベースLLMが直面する4つの根本的限界

1. 物理世界の理解の欠如

テキストのみで学習するため、物理法則や空間的直観を持たない。常識外れな誤答（「人が空を飛ぶ」など）を生成する。

2. 持続的なメモリの欠如

訓練後のパラメータは凍結され、経験を蓄積して成長できない。ワーキングメモリもコンテキストウィンドウに限定される。



3. 推論・論理能力の弱さ

統計的パターン学習に依存し、厳密な逐次推論ができない。複雑な論理や数理問題で不安定になり、「幻覚」を引き起こす。

4. 複雑な計画能力の欠如

長期的な目標達成のためのサブゴール設定や行動計画ができない。目的指向のエージェント機構を内部に持たない。

「理解」か、それとも精巧な「模倣」か

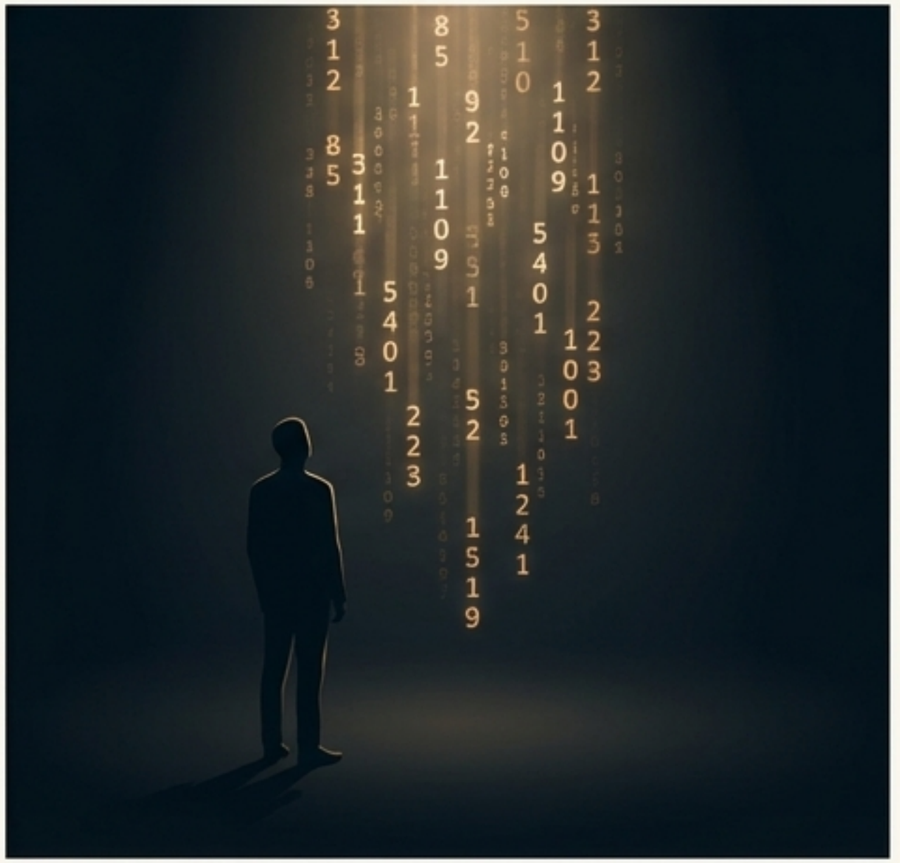
確率的オウム

「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」

Emily Bender, et al.

暗闇の中で数字の羅列を見る存在

「『怒り』という単語のトークン列は学習できても、怒りそのものを体験・理解しているわけではない。記号操作と意味理解は断絶している」



批判への反論：行き止まりか、進化の途中か

実用上の絶大な価値 (Immense Practical Value)

批判派も、LLMがAGIには到達しないとしても、その実用価値は極めて高いと認めている。

**Analogy:
「車のセールスマン」の例え。セールスマンは車を造れなくても、その知識には大きな価値がある。LLMも世界を創造できなくても、人間を支援する有益なツールである。

弱点を補う研究の進展 (Progress in Overcoming Weaknesses)

指摘されている欠点（知覚、記憶、推論）は、研究者たちが既に取り組んでいる課題でもある。

**Examples:
マルチモーダル化（GPT-4）、外部ツール連携（RAG）、思考プロセス支援（Chain-of-Thought）など、LLMの弱点を補う手法が次々と登場している。

**Conclusion: 総じて、「テキストLLM＝行き止まり」ではなく、進化の過程にある一形態と見る視点が重要である。

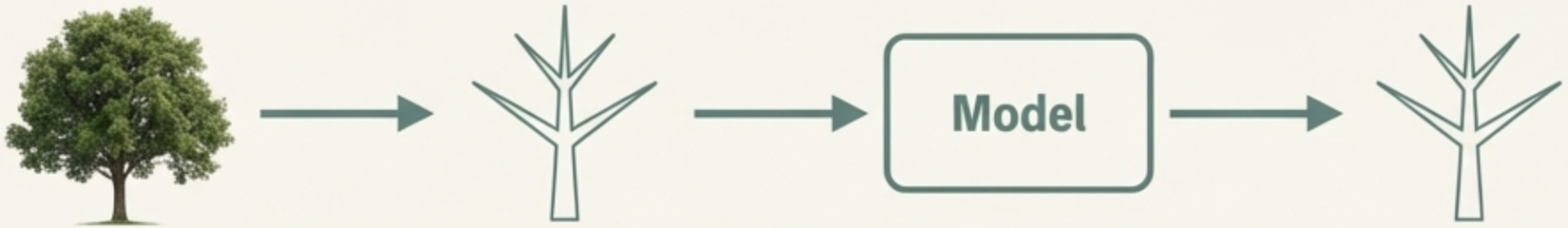
言葉の先へ：真のワールドモデルの構築

Core Concept: Joint Embedding Predictive Architecture（JEPA）

From: 旧来の生成モデル

「ピクセルを予測する」- 観測データそのものを再現しようとするため、木の葉の揺れなど不要な情報までモデル化しようとし、計算コストが高い。

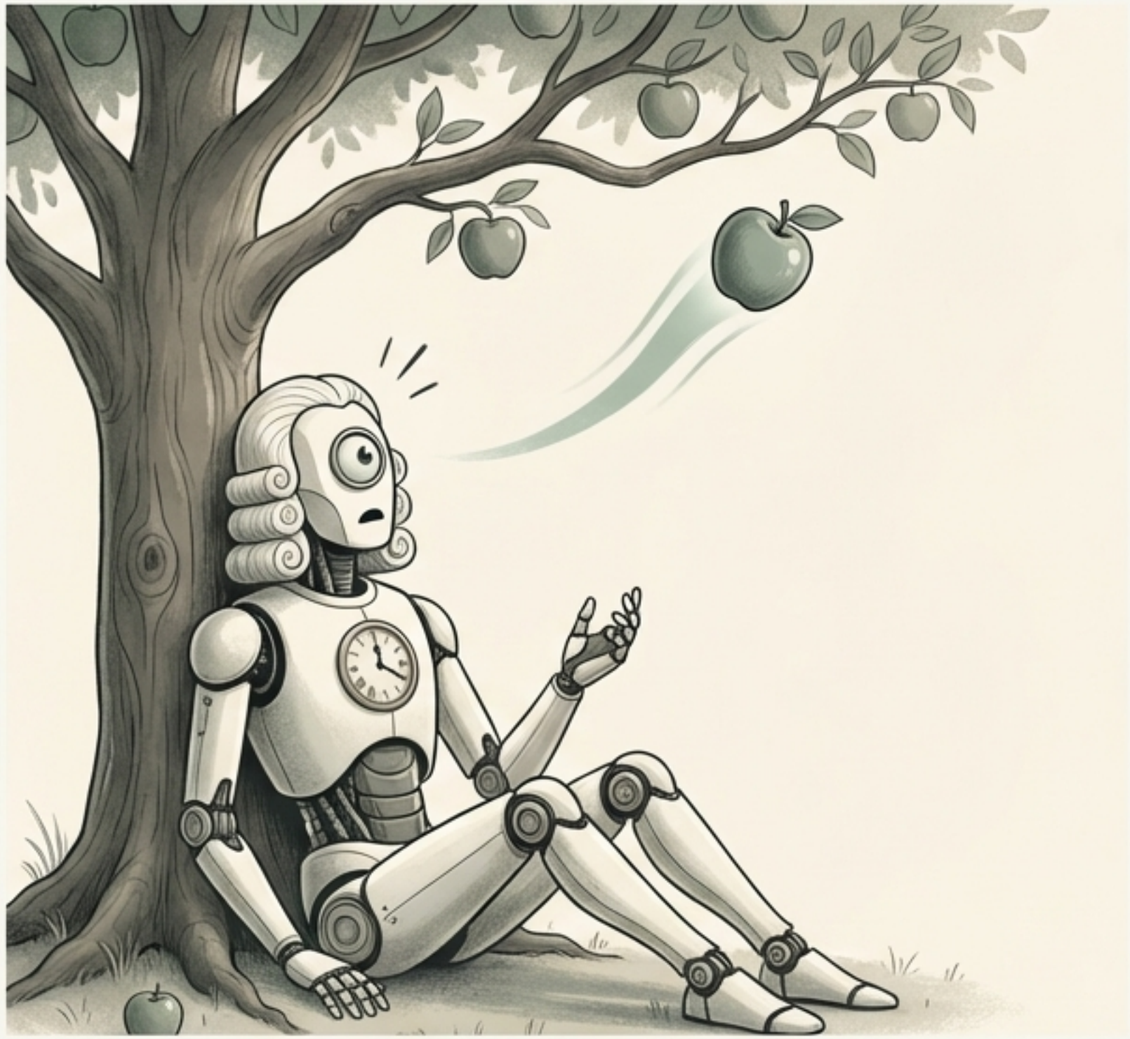
To: JEPA



「世界の抽象的な構造を予測する」- 生データを直接生成せず、その背後にある抽象的な表現（エンベディング）を予測。本質的な構造のみを捉え、効率的に世界のメンタルモデルを学習する。

Source: Yann LeCun (Meta) が提唱する次世代アーキテクチャ

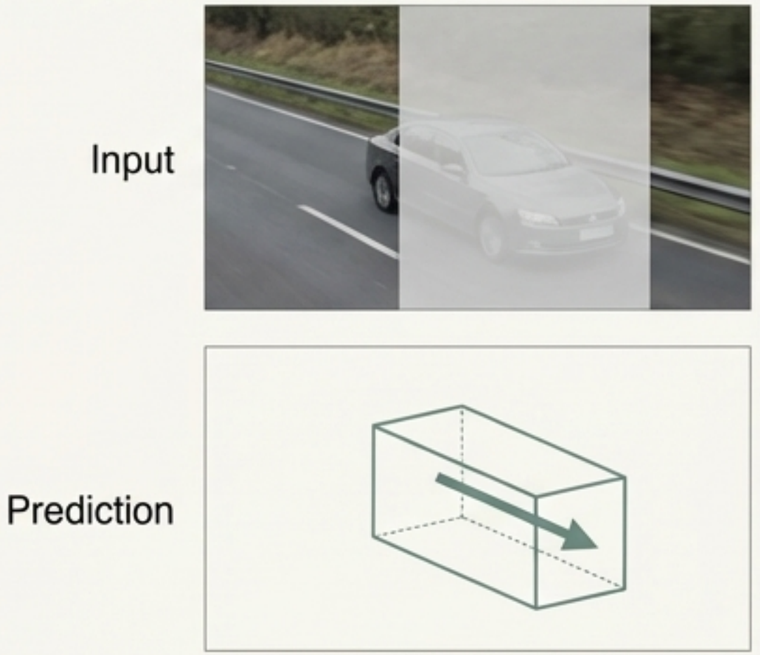
物理法則の直観を学習するAI



JEPAベースのモデルは、物理世界の振る舞いを学習し、常識に反する事象に対して「驚き」（高い予測誤差）を示す。

V-JEPA (Video JEPA) の実証

動画の一部を隠し（マスクし）、その部分で何が起きているかの抽象表現を予測するよう訓練。

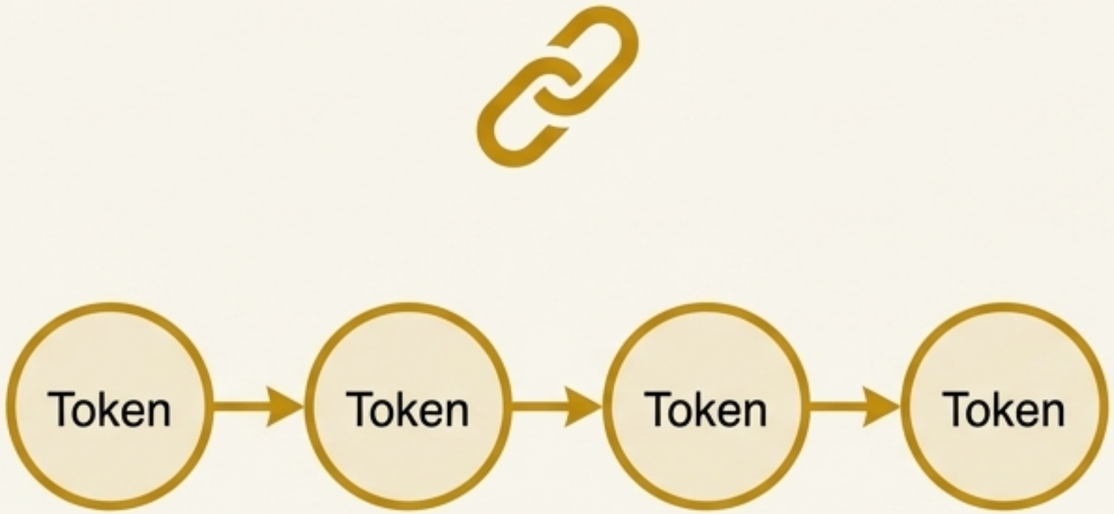


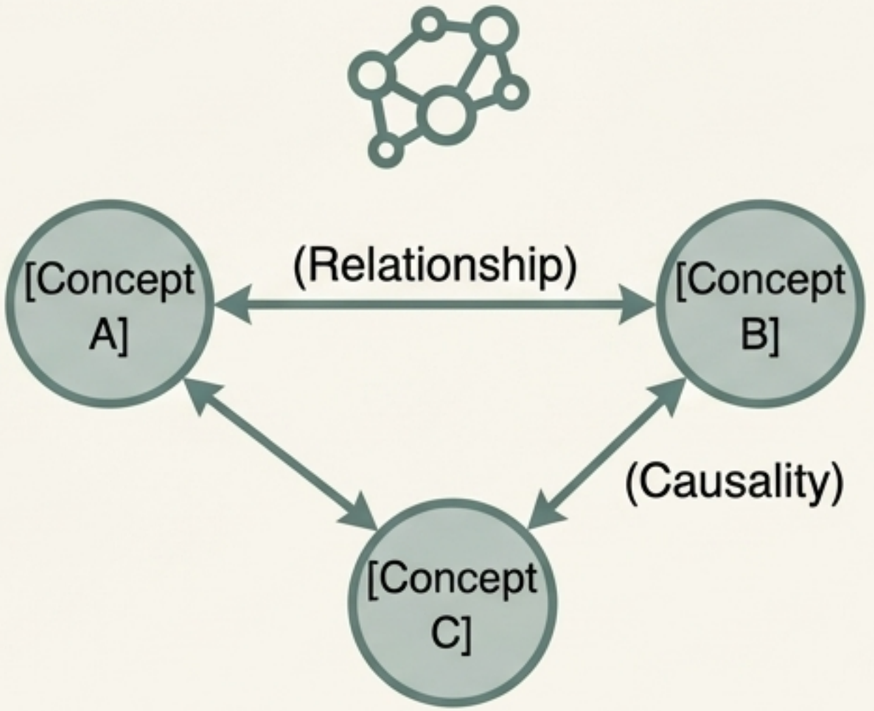
物体が突然消えたり、すり抜けたりする「あり得ない」映像を見せると、モデルは予測矛盾を示した。
これは、オブジェクトの永続性などの物理直観を内部に獲得したことを意味する。

単語の連鎖から、概念のネットワークへ：Large Concept Model（LCM）

Core Idea: 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」を操作対象とすることで、より構造的な推論を目指す。

Objective: 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を構築する。「知識と推論の融合」により、幻覚を低減し、信頼性を向上させる。





今後の展望：LLMの限界を克服するロードマップ（3-5年）

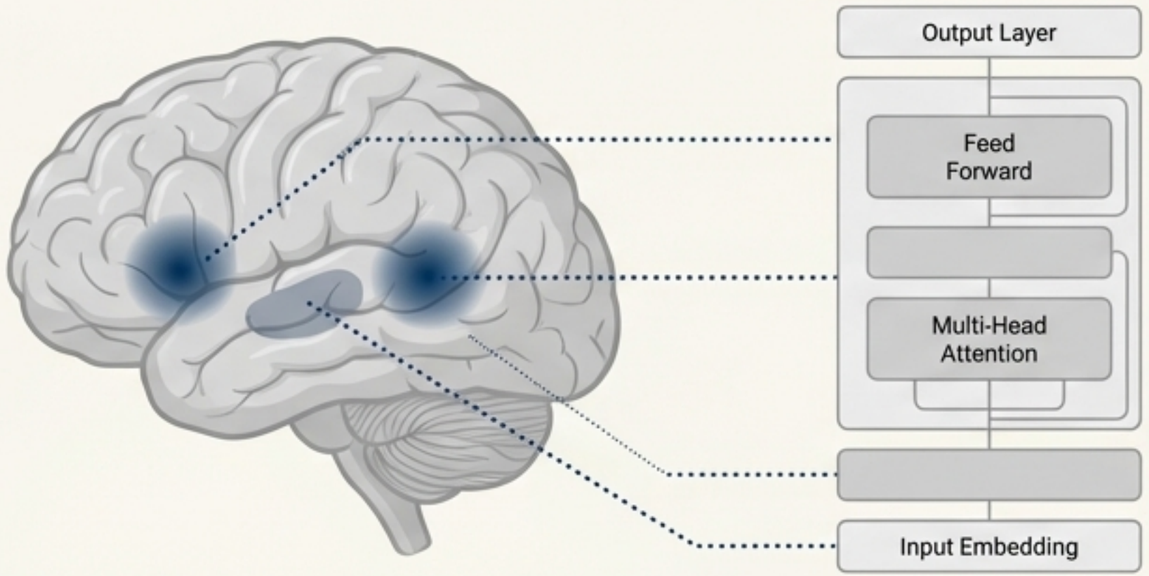
限界 (Limitation)	アプローチ (Approach)	具体的な技術 (Specific Technologies)	目指す能力 (Target Capability)
物理世界の理解不足	マルチモーダル学習 / 身体性	**JEPA**, V-JEPA, ロボティクスとの結合	物理的常識、因果理解
持続的なメモリの欠如	経験の蓄積と継続学習	**ReasoningBank**, 長期記憶フレームワーク	経験から学び成長するエージェント
推論・論理能力の弱さ	構造化推論と自己検証	**LCM**, CoT, 内部ツール利用, プロセス監督	信頼性と説明性の高い推論
複雑な計画能力の欠如	目的指向のエージェント化	階層型エージェント, 自己改善ループ（AutoGPT等）	自律的なタスク遂行、プランニング

脳との収斂：予測する機械としての共通原理

Core Theory: 脳科学の予測処理理論 (Predictive Processing) は、脳が常に外界を予測し、予測誤差を最小化する機械であると示唆する。

Evidence of Convergence: MITの研究：次単語予測精度が高いLLMほど、人間が文章を読む際の脳活動パターン（ブローカ野など）とよく合致する。

Insight: 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つつある可能性を示唆している。



次なる質的飛躍へ：言語の模倣から世界モデルの構築へ

LLMは、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した。しかし、その構造的限界も明らかになった。

現在の焦点は、「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」へと移行している。記憶し、考え、計画し、行動する総合知を目指す挑戦が始まった。

Yann LeCunが予測するように、今後3〜5年で新たなAIアーキテクチャが台頭し、現在の限界を克服するだろう。だろう。我々は、AIが再び質的飛躍を遂げる瞬間の入り口に立っている。その原動力こそが、「真のワールドワールドモデル」の導入である。