LLM における知識と社会的推論能力の進化的解析

原　匠[1]　　有田　隆也[1]　　鈴木　麗璽[1]
[1] 名古屋大学 大学院情報学研究科
Graduate School of Informatics, Nagoya University

概要

大規模言語モデル（LLM）は高度な認知機能を示すが，その能力の形成過程の理解は不十分である．本研究は，LLM を環境に適応する主体と捉え，遺伝的アルゴリズムを用いてゼロに近い能力状態から LoRA アダプタを進化させることで，認知タスクの性質が能力獲得過程に与える影響を分析した．知識集約型タスク（MMLU）と社会的推論タスク（ToMBench）を用いた実験の結果，環境の違いが適応度地形の形状に顕著な差をもたらすこと，進化経路は初期に重複し，後にタスク特異的に分岐すること，MMLU 適応個体は ToM タスクへ部分的に汎化するが逆は成立しないという非対称性があることが分かった．また，知識タスクへの適応がパターン認識的な社会的推論能力を向上させる一方で，深い推論能力とはトレードオフの関係にあることも示唆された．

1　はじめに

大規模言語モデル（LLM）の開発において，高度な認知能力の獲得メカニズムや，異なる能力間の相互関係（例えば知識と推論の関連性）は未解明な点が多い．既存の研究の多くは，学習済みモデルに対する評価や内部パラメータの静的な解析に留まっている．一方，生物学的な観点に基づき，環境への適応過程そのものを観察することで，機能の起源や構造を理解するアプローチがありうる．LLM の性能改善を目指した関連する取り組みとしては，進化的モデルマージ [1] 等の最適化手法が提案されているが，LLM の適応進化そのものの理解は不十分であると考える．

そこで，本研究は，進化計算に基づく LLM の最適化手法である GENOME[2] を応用し，LLM の認知能力の進化過程を分析することでその特徴を理解することを目的とする．具体的には，LLM のパラメータ（LoRA）を遺伝子と見なし，進化的アルゴリズムを用いてタスクに適応させる実験系を構築する．特に，幅広い知識を要する「MMLU」と，他者の心的状態を推論する「ToMBench」という性質の異なる 2 つのタスク環境において，ランダムな初期状態からどのように能力が獲得され，またそれらが互いにどのように汎化するかを比較分析する．これにより，LLM における「知識」と「社会的推論」の構成的な関係性と，能力獲得におけるトレードオフのメカニズムを明らかにする．

2　手法

2.1　GENOME[2] を用いた LLM の認知能力の進化

GENOME は，LLM の重みパラメータを遺伝子型と見なし，生物進化のメカニズムを模倣して最適化を行うフレームワークである．計算コストを抑制するため，巨大なベースモデルの重みは凍結し，追加学習用の軽量なパラメータ行列である LoRA (Low-Rank Adaptation)[3] のみを進化の対象とする．GENOME における進化のサイクルは以下の手順で行われる．

1. 適応度評価（ベンチマーク課題の正答率）
2. 選択 (ルーレット選択)
3. 交叉（TIES-Merging[4]：親のパラメータ間の干渉を抑制しながら特徴を合成）
4. 突然変異（子個体の重みに対し，一定の確率でガウスノイズを加算）

この評価から変異までのサイクルを規定世代数（G）繰り返すことで，集団はタスク環境に適応していく．詳細は文献 [5] を参照されたい．

2.2　初期集団の生成と適応過程の観測

オリジナルの GENOME は，特定のタスクにおける「性能の最大化」を主眼としている．そのため，

初期集団の生成には，既にタスク学習済みの複数の LoRA アダプタを読み込み，それらを混合することで，高い初期性能からのスタートを切る手法がとられている．

一方，本研究の目的は性能競争ではなく，LLM がタスク能力を獲得する「進化過程の理解」にある．そこで本研究では，初期集団の生成方法を変更し，意図的に適応能力を持たない集団を作成する．具体的には，LoRA アダプタの行列 A, B の全パラメータを正規分布 $N(0, \sigma^2)$ に従う乱数で初期化する．これにより，タスク性能が実質的にゼロの状態からスタートし，環境（タスク）からの選択圧のみによって，内部表現がどのように構造化され，能力が創発するかを観測することを可能にする．

2.3 認知タスクに基づく適応度評価

本研究では，性質の異なる 2 種類のベンチマークタスクにおける正答率を，各環境への適応度として定義する．一つ目の環境は知識集約型タスクの MMLU[6] である．これは STEM や人文科学を含む 57 の専門分野から構成され，モデルが持つ広範な世界知識を 4 択形式で評価するものである．二つ目の環境は社会的推論タスクの ToMBench[7] である．これは他者の感情や信念など 6 つのカテゴリを通じて「心の理論」を評価するもので，ストーリーの文脈理解を要する多肢選択問題である．各実験試行では，これらいずれかのタスクを評価関数として設定し，その正答率（0 から 1 の範囲）の最大化を目指して集団を進化させる．

2.4 遺伝子型の可視化

高次元なパラメータ空間における進化経路を追跡するため，変分オートエンコーダ（VAE）と UMAP を用いた次元圧縮手法を採用する．各レイヤーの LoRA 行列を結合したベクトルを VAE で潜在空間に圧縮し，さらに UMAP で 2 次元に投影することで，世代ごとの個体群の移動と，タスク間の遺伝子型の差異を可視化する．

3 実験と分析

3.1 設定

ベースモデルには Gemma-2-2b-it を用い，MMLU を適応度として用いた環境条件（MMLU 環境）と ToMBench を用いた環境（ToMBench 環境）それぞれについて 3 回の独立した進化実験（試行）を行った．主な設定を表 1 に示す．

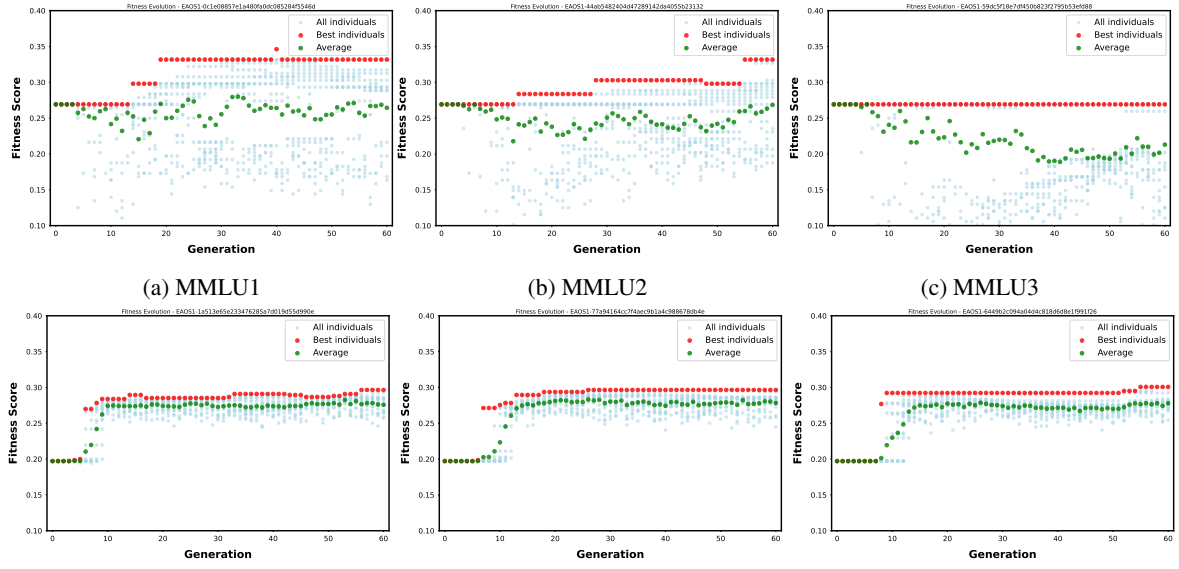
表 1: 実験パラメータと設定，各パラメータの詳細は文献 [2] 参照．

LLM	Gemma2-2b-it
最大世代数	G=60
初期集団のサイズ	P=10
LoRA アダプターのランク数	r=8
アダプター初期化時の正規分布の分散	$\sigma^2 = 0.02$
交叉率	cr=0.8
個体突然変異率	imr=0.15
遺伝子突然変異率	gmr=0.01
突然変異時の正規分布の分散	$\sigma^2 = 0.01$
エリート比率	$\alpha = 0.05$
VAE の潜在空間の次元	$Z \in \mathbb{R}^{1024}$

3.2 適応度の推移

図 1 に各環境における適応度の推移を示す．まず全体的な傾向として，両環境ともに初期のランダムな状態から進化に伴う性能向上が確認されたが，最終的な適応度の最大値はいずれも 0.32 程度に留まった．これはゼロからの学習としては有意な向上であるが，タスクを完全に習得するには至っていないことを示している．しかし，その上昇の過程には環境間で顕著な違いが見られた．MMLU 環境では，世代ごとの適応度の分散が大きく，高い個体と低い個体に二極化しつつ，特定の世代で急激な上昇が見られる階段状の進化を示した．一方，ToMBench 環境では分散が小さく，滑らかに最大適応度が上昇し，早期に収束する傾向が見られた．

適応環境による適応度推移の明確な差は，二つのタスクがパラメータ空間内に持つ適応度地形の構造的差異を示唆する．MMLU 環境における鋭利で分散の大きい推移は，知識タスクが険しい地形を持つことを意味する．これは，広範かつ正確な事実知識を再現するためには，パラメータ構成が特定の狭い領域に収束する必要があるためと考えられる．この場合，わずかな突然変異でもモデルの内部表現が破壊され，適応度が滑落するリスクが高い．対照的に，ToMBench 環境におけるなだらかで安定した推移は，社会的推論タスクが「台地状」の地形を持つことを示唆する．文脈理解や感情推測といった能力は，表面的な手がかりからのパターン認識で解答可



(a) MMLU1　(b) MMLU2　(c) MMLU3

(d) ToMBench1　(e) ToMBench2　(f) ToMBench3

図 1: それぞれの実験試行における適応度の分布と推移．各世代で記録された最大の適応度を示す赤い点は，個体を区別していない．

能であり，厳密なパラメータ調整を必要としない．そのため変異に対する頑健性が高く，柔軟な解法が許容されていると解釈できる．

3.3　遺伝子型の進化経路

VAE により圧縮された遺伝子型の UMAP プロット（図 2）を観察すると，進化の初期段階（第 1〜10 世代付近）では，両環境の個体群は近い領域を推移している．これは，タスクの特異性に関わらず，まず基本的な言語処理能力や出力形式への適応といった共通の基盤能力が獲得されることを示唆する．中盤以降，経路は明確に分岐し，環境ごとに特異的なクラスタを形成した．これは能力獲得が「一般的スキル」から「専門的スキル」へと階層的に進むことを裏付けている．

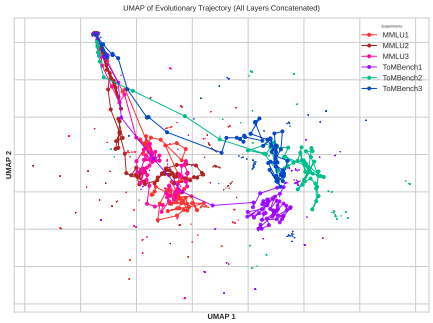
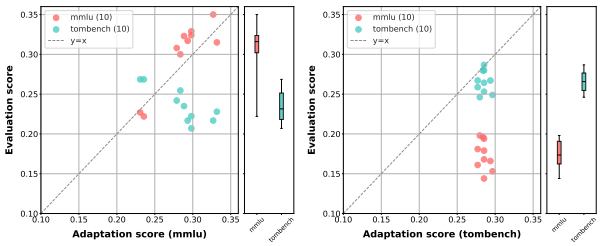


図 2: 世代数の増加に伴う遺伝子型の移動．左上の塊が進化初期，右下が後期．各世代の UMAP 上の点の重心をとり，代表点として線で結んでいる．

3.4　タスク間での汎化性能の比較

LLM が示す性質の異なる能力の間に何らかの特徴的な関係が存在するかを検討するために，一方の環境に適応した個体群が，もう一方の未知の環境に対してどれほどの適応度を示すことができるかを評価した．（図 3）グラフの横軸は進化環境における適応度，縦軸は評価タスクのスコアを表す．このプロットにおいて，右上がりの分布（正の相関）が確認されれば，適応タスクでの能力向上が未知のタスクへ汎化していることを意味する．逆に，無相関あるいは右下がりの分布（負の相関）であれば，その能力適応が他方のタスク性能に寄与しない，あるいはトレードオフの関係にあることを示唆する．

MMLU 環境で進化した個体は，ToMBench タスクにおいても初期状態より高いスコアを記録し，正の相関が見られた．しかし，ToMBench 環境で進化した個体は，MMLU のスコアが低く，一部では負の相関が見られた．この非対称な汎化は，知識集約型タスクへの適応が，文脈理解やパターン認識等の社会的推論に必要な基礎能力の一部を副次的に強化する一方で，社会的推論タスクへの特化は，広範な知識ベースを必要としないため，知識タスクへの応用が効かないことを示していると考えられる．



(a) MMLU2
-mmlu: 0.905
-tombench: -0.821

(b) ToMBench2
-mmlu: 0.388
-tombench: -0.052

図 3: それぞれの実験試行における最終世代の適応タスクのスコアと評価タスクのスコアの関係．タイトル下の数字は Peason の相関係数である．

3.5 ToM カテゴリ別のトレードオフ

MMLU 適応個体が ToM タスクでどのような振る舞いをするか詳細に分析ため，ToMBench の 6 カテゴリごとのスコア変動に注目した．クロス実験において顕著な負の相関を示した MMLU2 について，結果を図 4 に示す．「Emotion（感情推測）」や「NLC（皮肉等の理解）」，「Desire（欲求）」のスコアは上昇傾向にあったが，「Belief（信念）」や「Knowledge（知識状態の推論）」，「Intention（意図）」は変化がないか，むしろ低下していた．

Emotion や NLC は，一般的な知識を用いて文脈のパターンから特定の感情描写や皮肉をイメージすることが容易である可能性が報告されている [7]．対して Belief 課題は，「他者が誤った信念を持っている」という状況をシミュレートする必要があり，高度な論理的整合性が求められる．このことがカテゴリごとに異なる傾向の要因になったと考えられる．

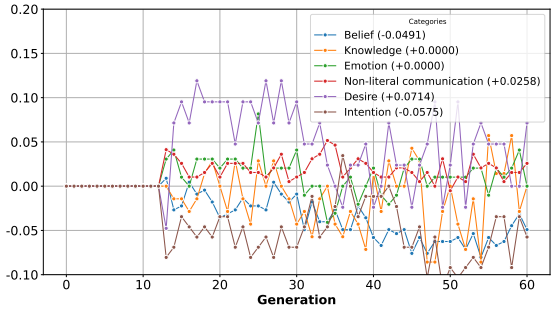


図 4: MMLU2 における，ToMBench の 6 カテゴリを区別した適応度の推移．初期の個体のスコアを 0.00 とした適応度の相対値．

4 考察

実験結果から，タスクの性質が LLM の進化に大きな影響を与えることが明らかになった．特に MMLU への適応過程において，モデルは膨大な知識問題に正答するために，問題文のパターンから効率的に答えを導き出すヒューリスティクスを発達させたと考える．この能力は，同様にパターン認識で解ける ToM タスク（Emotion 等）には転移したが，複雑な内部シミュレーションを要するタスク（Belief 等）においては，リソースの競合やショートカットへの過学習により，むしろ性能を阻害した可能性がある．これを適応コストの観点から解釈すれば，進化プロセスはコストの低い能力であるパターン認識を優先的に獲得し，高コストな能力である深い推論を後回しにする，あるいは犠牲にするというメカニズムが働いていると考えられる．

5 おわりに

本研究では，タスクの性質が LLM の能力獲得過程にいかなる影響を与えるかを解明するために，知識集約型タスク（MMLU）と社会的推論タスク（ToMBench）という異なる環境において，遺伝的アルゴリズムを用いてゼロから LoRA アダプタを進化させる比較実験を行った．

実験の結果，タスクの性質が適応度地形の形状（険しさ）に顕著な差異をもたらすこと，および進化経路が初期の重複を経てタスク特異的に分岐することが明らかになった．特に，知識タスクに適応した個体は社会的推論タスクへ部分的に汎化するが，その逆は成立しないという非対称性が確認された．

この結果は，広範な知識への適応が浅い社会的推論能力を底上げする一方で，深い推論能力の獲得とはトレードオフになる可能性を示唆している．本研究で得られた知見は，LLM の能力が単に独立して存在するのではなく，相互に干渉し合いながら階層的に構築されることを実証しており，今後のモデル設計やカリキュラム学習に対して重要な示唆を与えるものである．

一方で，本研究は小規模なモデルと限定的なタスクを用いた初期検証に留まっている．今後は，より大規模なモデルや多様な認知タスクを用いた実験を行い，能力獲得のメカニズムをより包括的に解明することが課題である．

参考文献

[1] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. **Nature Machine Intelligence**, Vol. 7, No. 2, pp. 195–204, 2025.

[2] Yiqun Zhang, Peng Ye, Xiaocui Yang, Shi Feng, Shufei Zhang, Lei Bai, Wanli Ouyang, and Shuyue Hu. Nature-inspired population-based evolution of large language models. **arXiv preprint arXiv:2503.01155**, 2025.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[4] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. **Advances in Neural Information Processing Systems**, Vol. 36, pp. 7093–7115, 2023.

[5] 原匠, 有田隆也, 鈴木麗璽. LLM の適応進化においてタスクの性質が進化過程に与える影響. 人工知能学会第二種研究会資料, Vol. 2025, No. ALIFE-010, 06, 2025.

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. **arXiv preprint arXiv:2009.03300**, 2020.

[7] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. ToMBench: Benchmarking theory of mind in large language models. **arXiv preprint arXiv:2402.15052**, 2024.