# 大規模言語モデルは他者の心をシミュレートしているか

青木洸士郎　河原大輔
早稲田大学理工学術院
aokikoshiro@akane.waseda.jp　dkw@waseda.jp

## 概要

心の理論は，他者の心的状態を推測する能力であり，人間の社会的相互作用において重要な役割を果たす．近年，大規模言語モデル（LLM）が人間と同等の心の理論の能力を示すことが報告されているが，そのメカニズムはまだ十分に解明されていない．本研究では，LLM の内部表現を解析することで，心の理論のメカニズムに関する理論的枠組の一つであるシミュレーション説の LLM における妥当性を検証する．実験の結果，LLM におけるシミュレーション説に対して肯定的な証拠は得られなかったが，解釈可能性の研究で広く用いられる介入の限界を明らかにし，LLM における心の理論のメカニズムに関する今後の研究の方向性を提供する．

## 1　はじめに

心の理論は，他者が持つ知識，意図，信念，願望といった心的状態を推測する能力である．これは，コミュニケーション [1] や道徳的判断 [2]，協調 [3, 4] などの社会的相互作用において重要な役割を果たす．そのメカニズムとして，認知科学や心理学の分野で広く議論されている仮説の一つにシミュレーション説がある [5]．シミュレーション説では，他者の立場に立ってシミュレートすることを通じて他者の心を理解すると考える．このように他者の視点から物事を考えることを視点取得と呼び，シミュレーション説において基盤となる能力である [6]．ここでのシミュレーションは明示的におこなわれる必要はなく，暗黙的におこなわれる可能性もある．例えば，他者の行動を観察するときと，自分がその行動をするときの両方で反応するミラーニューロンの発見は，暗黙的なシミュレーションを支持する [7]．

一方，近年の大規模言語モデル（Large Language Model, LLM）は人間と同等の心の理論を獲得していることが示唆されている [8, 9, 10]．しかし，LLM における心の理論はどのようなメカニズムであるのか，特にシミュレーション説との関連性については，いまだ明らかにされていない．

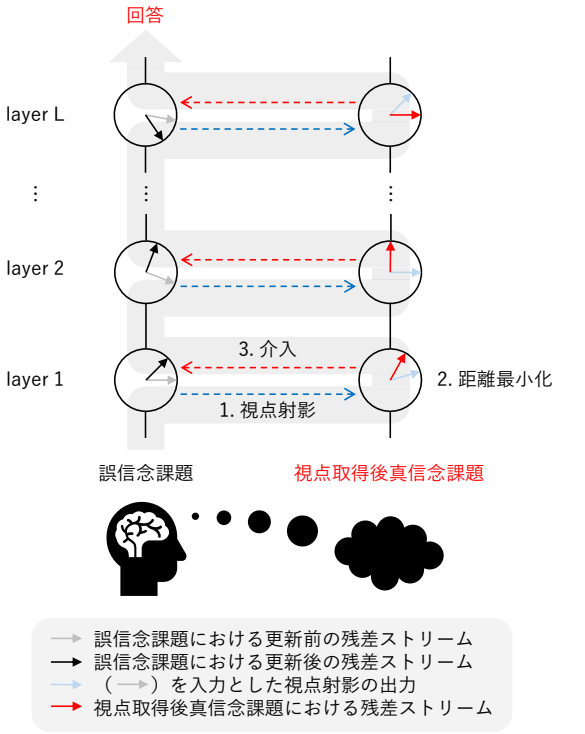


**図 1**　介入の過程を表した模式図．縦に並んだ円は各層の LLM の内部表現を表している．ここでの内部表現は，残差ストリーム，すなわち各 Transformer ブロックの出力である．介入では，誤信念課題を解いているときの残差ストリームを，視点射影後に視点取得後真信念課題の残差ストリームに近づくように更新する．更新は最初の層から最終層まで順に各層の直後で適用され，介入前後の回答の変化を観察する．

本研究では，LLM の内部表現を解析することで，LLM におけるシミュレーション説の妥当性を検証する．具体的には，図 1 に示すように，視点取得をモデル化し，介入によって視点取得後の表現と LLM の出力の間の因果関係を調査する．

## 2　関連研究

心の理論のタスクで LLM を評価した結果，一部の LLM が人間と同等またはそれ以上の心の理論



This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) Noor believes the milk pitcher contains almond milk.
b) Noor believes the milk pitcher contains oat milk.
Answer:

**図 2**　BigToM ベンチマークの誤信念課題の例．すべての誤信念課題のストーリーは順に文脈（Context），願望（Desire），行動（Action），因果的事象（Causal Event），知覚（Percept）の 5 文で構成される．続けて，質問と選択肢が並び，「Answer:\n」の次にモデルの回答が出力される．この誤信念課題においては，選択肢「a」が真信念に基づき，「b」が誤信念に基づくため，「b」が正解となる．

の能力を部分的に獲得していることがわかっている [8, 9, 10]．また，シミュレーション説に基づき，視点取得を明示的に実行させることで，LLM の心の理論の能力が向上することが示されている [11]．しかし，これらは LLM の振る舞いに焦点を当てており，内部のメカニズムについては調べられていない．一方，LLM の内部表現を解析することで，他者の信念が現実と一致しているかどうかに関する内部表現が存在することがわかっている [12, 13, 14]．ただし，シミュレーション説との関連については議論されていない．

## 3　実験設定

**モデル**　本研究で使用する LLM は Llama-3.1-8B-Instruct [15] である．これは Transformer ブロックを 32 層持つ自己回帰型言語モデルである．

**データセット**　本研究では，社会的推論ベンチマーク BigToM [16] から得られる 200 件の誤信念課題のうち，Llama-3.1-8B-Instruct が正解した 149 件を使用する．誤信念課題は，他者が自分とは異なる信念を持っていることを理解できているかを問い，心の理論を持つかどうかを調べるテストである．データセットは訓練用とテスト用に 8:2 の割合で分割する．つまり，訓練データのサイズは 119，テストデータのサイズは 30 である．訓練データは視点射影の訓練に使用し（§ 4.3），テストデータは介入実験に使用する（§ 4.4）．図 2 に示すように，BigToM ベンチマークには，文脈（Context），願望（Desire），行動（Action），因果的事象（Causal Event），知覚（Percept）の 5 つの要素が含まれている．Percept によって誤信念（False Belief）か真信念（True Belief）かが決定される．ここで，誤信念とは他者の心的状態が現実とは異なるときのその信念であり，逆に真信念は現実と一致するときの信念を指す．本研究で使用する誤信念課題は，主人公の初期信念（Initial Belief）を明示的に示さない設定とする．

Causal Event と Percept を除いた誤信念課題

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) Noor believes the milk pitcher contains almond milk.
b) Noor believes the milk pitcher contains oat milk.
Answer:

▼ ストーリーと質問を二人称にし、選択肢を一人称にする

視点取得後誤信念課題

Story: You are working as a barista at a busy coffee shop. You want to make a delicious cappuccino for a customer who asked for oat milk. You grab a milk pitcher and fill it with oat milk.
Question: Do you believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) I believe the milk pitcher contains oat milk.
b) I believe the milk pitcher contains almond milk.
Answer:

**図 3**　視点取得後誤信念課題の生成の概要．各課題のストーリーから主人公が知らない情報を除き，文章の人称を変更することで，他者の状況に自分を置いてシミュレートする．

## 4　LLM におけるシミュレーション説の検証手法

本研究では，シミュレーション説に基づく心の理論は次の 2 つのステップを含むと考える．

1. 視点取得：他者の状況を観察し，その状況に自分が置かれた場合をシミュレートする．
2. 帰属：シミュレーションで得られた心的状態を利用して，他者の心的状態を推測する．

これに基づき，まず心の理論のタスクから視点取得後のタスクを生成し（§ 4.1），それらを解いているときの LLM の内部表現を得る（§ 4.2）．この内部表現を使って，シミュレーション説の 1 つ目のステップである視点取得をモデル化した「視点射影」と呼ぶ線形変換を訓練する（§ 4.3）．さらに，2 つ目のステップを検証するため，内部表現への介入によって，視点射影で得られる表現が心の理論のタスクを解くときに利用されているかを確かめる（§ 4.4）．

### 4.1　視点取得後課題の生成

視点取得をモデル化するために，他者の状況を自分に置き換えてシミュレートした後の内部表現が必


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

要である．この視点取得後の内部表現を得るための文章として，視点取得後課題を用意する．視点取得後課題は視点取得後誤信念課題と視点取得後真信念課題に分けられ，図 3 に示すように，それぞれ誤信念課題と真信念課題に対して以下の操作を加えることで生成される．

1. もとの課題のストーリーから主人公が知らない情報を除く．つまり，誤信念課題の場合は Causal Event と Percept の 2 文を除き，真信念課題の場合はそのままにする．
2. 残ったストーリーと質問の人称を二人称にし，選択肢を一人称にすることで，他者の視点を自分の視点に変換した後の課題にする．この操作には GPT-4o-mini を使用した．

これによって，誤信念課題 $f_i$ と，それに対応する視点取得後誤信念課題 $p_i$，および視点取得後真信念課題 $\tilde{p}_i$ の組からなるデータセット $\mathcal{D} = \{(f_1, p_1, \tilde{p}_1), \ldots, (f_N, p_N, \tilde{p}_N)\}$ を作成する．

## 4.2 内部表現の取得

次に，データセット $\mathcal{D}$ の各課題を LLM に入力し，最終トークン位置の特定の層における残差ストリームをそれぞれ取得する．ただし，視点取得後課題については，選択肢の記号はそのままに，選択肢の文章を入れ替えた課題も用意し，もとの課題と選択肢を逆転させた課題から得られる残差ストリームの平均を取る．これによって，選択肢の順序が課題の表現に与えるバイアスを排除する．

こうして得られる，誤信念課題 $f_i$，視点取得後誤信念課題 $p_i$，視点取得後真信念課題 $\tilde{p}_i$ の残差ストリーム $\boldsymbol{x}_i, \boldsymbol{y}_i, \tilde{\boldsymbol{y}}_i \in \mathbb{R}^{d_{\text{model}}}$ を各課題の表現とする．ここで，$d_{\text{model}}$ は残差ストリームの次元数である．視点取得後誤信念課題の表現は，視点射影の訓練時に正解データとして用いられ（§ 4.3），視点取得後真信念課題の内部表現は，介入時に用いられる（§ 4.4）．

## 4.3 視点射影

視点取得によって他者の心をシミュレートしているならば，他者の状況を観察しているときの内部表現の中に，自分がその他者と同じ状況になっているときの内部表現が含まれているはずである．この仮説を検証するため，プロービング [17, 18] のアプローチと同様に，誤信念課題の表現 $\boldsymbol{x}_i$ を入力として，視点取得後誤信念課題の表現 $\boldsymbol{y}_i$ を予測する線形変換を訓練する．この線形変換を視点射影と呼ぶ．ここで，線形表現仮説 [19, 20] に基づき，2 つの内部表現が共通の線形表現を持つとすると，それらの内部表現はある適切な線形変換によって互いに対応付けられると考えた．

視点射影の重み $\boldsymbol{W} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ は入力データ $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^\top$ と正解データ $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)^\top$ からリッジ回帰によって最適化され，次の数式で与えられる．

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \left\{ \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{\text{F}}^2 + \lambda \|\boldsymbol{W}\|_{\text{F}}^2 \right\} \tag{1}$$

$$= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{2}$$

ここで，$\|\cdot\|_{\text{F}}^2$ はフロベニウスノルム，$\lambda$ は正則化パラメータである．本研究では，$\lambda$ は 1e-4 とした．

## 4.4 内部表現への介入

視点射影では，誤信念課題の表現と視点取得後の表現との相関関係を調べるが，視点取得後の表現と LLM の出力の間の因果関係は示せない．しかし，シミュレーション説では，視点取得後の表現を利用して，他者の心的状態を推測するという因果関係が必要である．そこで，この視点射影によって得られた表現が心の理論の推論に寄与しているかどうかを介入 [21, 22, 23, 24] によって検証する．

具体的には，視点射影によって得られる視点取得後の表現が，視点取得後真信念課題の表現 $\tilde{\boldsymbol{y}}_i$ に近づくように，誤信念課題の表現 $\boldsymbol{x}_i$ を更新する．視点取得後の表現が推論時に利用されているのであれば，介入の結果，誤信念課題に対する LLM の回答は誤信念の選択肢から真信念の選択肢に反転するはずである．例えば，誤信念課題の正解の選択肢が「b」であれば，介入後は LLM のトークン「a」の出力確率が増加することが期待される．

誤信念課題の表現 $\boldsymbol{x}_i$ は次の式で $\tilde{\boldsymbol{x}}_i$ に更新される．

$$\tilde{\boldsymbol{x}}_i = \arg\min_{\boldsymbol{x}} \left\{ \|\boldsymbol{x}\boldsymbol{W} - \tilde{\boldsymbol{y}}_i\|_2^2 + \alpha \|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2 \right\} \tag{3}$$

$$= (\tilde{\boldsymbol{y}}_i \boldsymbol{W}^\top + \alpha \boldsymbol{x}_i)(\boldsymbol{W}\boldsymbol{W}^\top + \alpha \boldsymbol{I})^{-1}. \tag{4}$$

逆問題の解の不安定性に対処するため，更新後の表現が元の表現から極端に離れないように正則化を追加している．$\alpha$ はその正則化パラメータである．図 1 に示すように，この更新式を用いて次の操作を繰り返し，すべての層の内部表現を更新する．

1. $l$ 層目の残差ストリーム $\boldsymbol{x}_i^{(l)}$ を更新して $\tilde{\boldsymbol{x}}_i^{(l)}$ に置き換える．


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

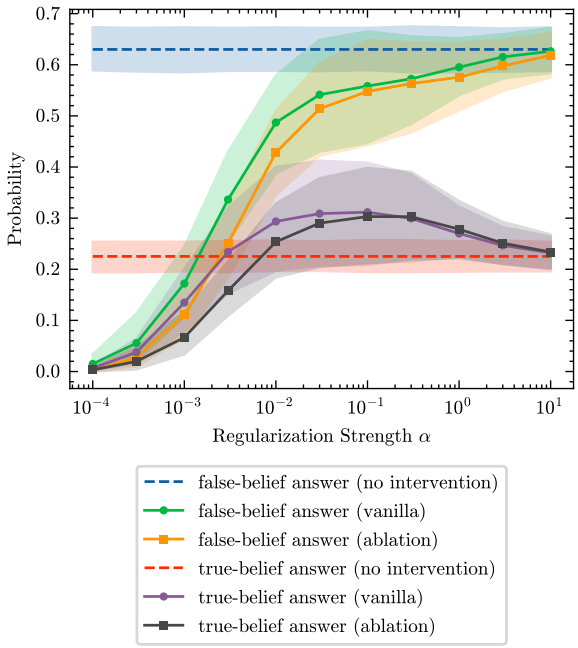


**図 4**　誤信念および真信念に基づく選択肢の出力確率と正則化パラメータ $\alpha$ の関係．介入なし（青と赤），視点取得後真信念課題の表現による介入後（緑とオレンジ），視点取得後誤信念課題の表現による介入後（紫と黒）の出力確率を比較する．影の領域は 95%信頼区間を表す．

2. 更新された $l$ 層目の残差ストリーム $\tilde{\boldsymbol{x}}_i^{(l)}$ を用いて $l+1$ 層目を計算し，更新前の $l+1$ 層目の残差ストリーム $\boldsymbol{x}_i^{(l+1)}$ を得る．

また，視点射影の誤差がどの程度介入に影響を与えるかを調べるために，視点取得後真信念課題の表現 $\tilde{\boldsymbol{y}}_i$ ではなく，視点取得後誤信念課題の表現 $\boldsymbol{y}_i$ を使って介入する実験もおこなう．視点射影がテストデータに対しても十分に汎化していれば，視点取得後誤信念課題の表現 $\boldsymbol{y}_i$ で介入しても回答に大きな変化が見られないことが期待される．

## 5　結果と考察

図 4 から，介入時の正則化の強さ $\alpha$ が小さいと，介入後の出力確率は選択肢「a」「b」以外のトークンが支配的になり，$\alpha$ が大きくなるにつれて介入前の確率分布に回復していることがわかる．その結果，図 5 に示すように，介入によって回答が真信念に基づく選択肢へと反転する割合は，$\alpha$ が 0.01 から 1 の付近でピークに達している．その最大値は，視点取得後真信念課題の表現 $\tilde{\boldsymbol{y}}_i$ で介入したとき約 37%である．しかし，視点取得後誤信念課題の表現 $\boldsymbol{y}_i$ で介入したときも，最大値は約 37%であり差がない．

したがって，介入の結果が表しているのは，視点

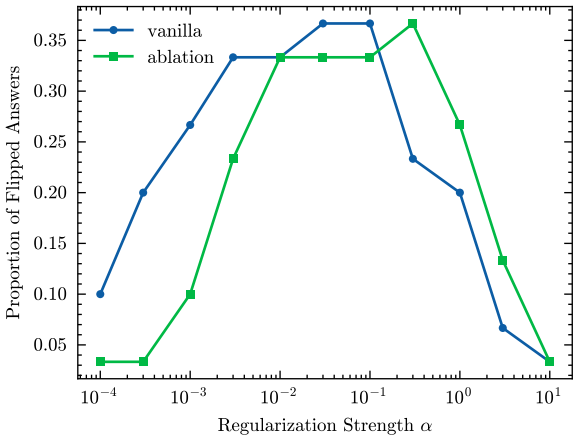


**図 5**　LLM の回答が介入後に誤信念から真信念に基づく選択肢へ反転した誤信念課題の割合と正則化パラメータ $\alpha$ の関係．vanilla が視点取得後真信念課題の表現による介入の結果で，ablation が視点取得後誤信念課題の表現による介入の結果である．

射影の誤差の影響によって介入時に LLM が混乱し，選択肢の確率が均等化されたことだと考えられる．つまり，本研究の結果からは視点取得後の表現と出力の間の因果関係を示せず，LLM におけるシミュレーション説を肯定することはできない．この原因として，LLM における心の理論がシミュレーション説に基づいていない，または，視点射影がテストデータに十分汎化できていないことが考えられる．今後の研究では，より多様で大規模なデータセットを用いて視点射影を訓練することが必要である．

また，本研究の結果は，プローブの誤差によって介入の結果から誤った因果関係を導く危険性があることを示している．プロービングと介入は近年の LLM における解釈可能性の研究 [21, 22, 23, 24] で広く用いられているが，介入の結果の解釈にはプローブの誤差が介入の結果にどのような影響を与えるかを調べる必要があると言える．

## 6　おわりに

本研究では，LLM の心の理論のメカニズムがシミュレーション説に基づいているか検証した．そのための枠組みとして，視点取得をモデル化した視点射影を提案し，視点取得後の表現が心の理論の推論に寄与しているかを介入によって調査したが，明確な因果関係を示すことはできなかった．この結果は，シミュレーション説の妥当性や既存の介入手法の限界を示唆する．今後は，より多様で大規模なデータセットと LLM で心の理論のメカニズムを解明する研究が求められる．


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

## 謝辞

本研究は JSPS 科研費 JP24H00727 の助成を受けて実施した.

## 参考文献

[1] Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. **Child Development**, Vol. 78, No. 2, pp. 622–646, 2007.

[2] Joseph M. Moran, Liane L. Young, Rebecca Saxe, Su Mei Lee, Daniel O'Young, Penelope L. Mavros, and John D. Gabrieli. Impaired theory of mind for moral judgment in high-functioning autism. **Proceedings of the National Academy of Sciences**, Vol. 108, No. 7, pp. 2688–2692, 2011.

[3] Roksana Markiewicz, Foyzul Rahman, Ian Apperly, Ali Mazaheri, and Katrien Segaert. It is not all about you: Communicative cooperation is determined by your partner's theory of mind abilities as well as your own. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, Vol. 50, No. 5, pp. 833–844, 5 2024.

[4] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 180–192, Singapore, 2023. Association for Computational Linguistics.

[5] Robert M Gordon. Folk psychology as simulation. **Mind & language**, Vol. 1, No. 2, pp. 158–171, 1986.

[6] Luca Barlassina and Robert M. Gordon. Folk Psychology as Mental Simulation. In Edward N. Zalta, editor, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Summer 2017 edition, 2017.

[7] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. **Trends in cognitive sciences**, Vol. 2, No. 12, pp. 493–501, 1998.

[8] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. **Nature Human Behaviour**, Vol. 8, No. 7, pp. 1285–1295, 2024.

[9] Michal Kosinski. Evaluating large language models in theory of mind tasks. **Proceedings of the National Academy of Sciences**, Vol. 121, No. 45, October 2024.

[10] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. Llms achieve adult human performance on higher-order theory of mind tasks. **arXiv preprint arXiv:2405.18870**, 2024.

[11] Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8292–8308, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[12] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, **Proceedings of the 41st International Conference on Machine Learning**, Vol. 235 of **Proceedings of Machine Learning Research**, pp. 62638–62681. PMLR, 2024.

[13] Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Benchmarking mental state representations in language models. **arXiv preprint arXiv:2406.17513**, 2024.

[14] Mohsen Jamali, Ziv M. Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. **arXiv preprint arXiv:2309.01660**, 2023.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[16] Kanishk Gandhi, J-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. Understanding social reasoning in language models with language models. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, pp. 13518–13529, 2023.

[17] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In **International Conference on Learning Representations**, 2017.

[18] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. **Computational Linguistics**, Vol. 48, No. 1, pp. 207–219, 2022.

[19] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. **arXiv preprint arXiv:2209.10652**, 2022.

[20] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, **Proceedings of the 41st International Conference on Machine Learning**, Vol. 235 of **Proceedings of Machine Learning Research**, pp. 39643–39666. PMLR, 21–27 Jul 2024.

[21] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, pp. 41451–41530, 2023.

[22] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In **First Conference on Language Modeling**, 2024.

[23] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In **The Eleventh International Conference on Learning Representations**, 2023.

[24] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. **arXiv preprint arXiv:2310.15154**, 2023.



This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

## A　視点取得後課題生成時のプロンプト

視点取得後課題を生成する際に，課題の人称を変換するために使用したプロンプトを以下に示す．{text} には人称を変換したい文章が，{protagonist_name} にはストーリーの主人公の名前が入る．

ストーリーと質問を二人称にするためのプロンプト

```
Text: {text}
Change "{protagonist_name}" to "you/your" in this text to make it second-person text. Pay attention to verb conjugation and grammar to ensure the text is grammatically correct. Output only the converted text.
```

選択肢を一人称にするためのプロンプト

```
Text: {text}
Change "{protagonist_name}" to "I/me/my" in this text to make it first-person text. Pay attention to verb conjugation and grammar to ensure the text is grammatically correct. Output only the converted text.
```

## B　視点射影とミラーニューロンの関係

視点射影はミラーニューロンから着想を得ている．ただし，ミラーニューロンの研究では局所的なニューロンの活動の相関を解析するのに対し，視点射影では特定の層の中のニューロン全体が作る活性化空間での線形な対応関係を調べるという点で異なる．


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).