Let's Put Ourselves in Sally's Shoes: 他人の靴プレフィリングは大規模言語モデルの心の理論を改善する

篠田一聡　北条伸克　西田京介　山﨑善啓　鈴木啓太　杉山弘晃　齋藤邦子
NTT 株式会社　人間情報研究所
kazutoshi.shinoda@ntt.com

概要

大規模言語モデル（LLM）の心の理論（ToM）を改善するために，新しい推論手法である 他人の靴プレフィリング（Shoes-of-Others prefilling）を提案する．他人の靴プレフィリングは，LLM の出力の冒頭を "Let's put ourselves in A's shoes."（A の立場に立って考えよう）で指定してその続きを生成させる単純な手法であり，広範な問題設定に適用可能である．対話及び物語を入力とする 2 つのベンチマークを用いて ToM を評価した結果，5 種類の心的状態，特に誤信念の理解において，一貫して提案手法による ToM 性能の向上が確認された．分析の結果，提案手法は予測を忠実に説明する思考の生成を促進することで ToM 性能が改善することが示唆された．

1　はじめに

心の理論（Theory of Mind, ToM）とは，他者の信念，意図，願望などの心的状態を推論する能力を指す [1]．ToM は，人間が他者と効果的に相互作用するために不可欠であると考えられている [2]．大規模言語モデル（LLM）は，メールの自動補完 [3]，共感的対話 [4]，説得 [5] といった，ToM を必要とする場面で利用されることが増えており，それに伴って LLM における ToM の能力も重要性を増している [6]．一方で，近年提案されたベンチマークにおいて，LLM の ToM 性能は依然として人間に及ばないことが，これまでの研究により示されている [6, 7, 8, 9, 10, 11, 12]．ToM データセットを用いた LLM の fine-tuning は，分布内汎化性能を向上させる一方で [9, 11]，過学習を引き起こし，分布外汎化性能を低下させることが報告されている [13]．

このような背景から，モデルパラメータの調整を必要とせず，設計上過学習を引き起こさない ToM のための推論手法に対する関心が高まっている

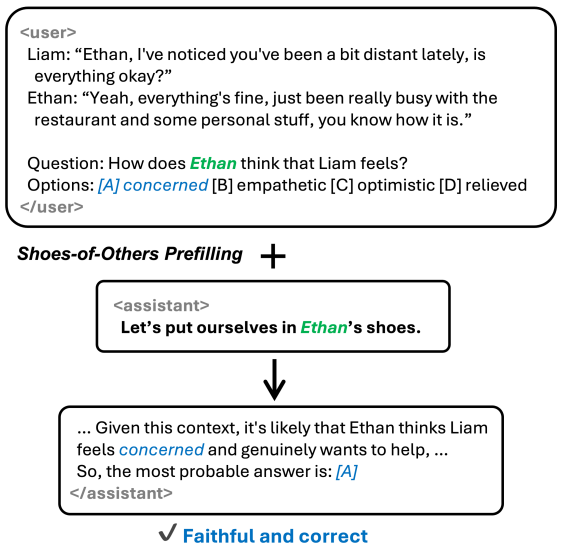


図 1　他人の靴プレフィリングの概要．この ToMATO [12] の例は，他人の靴プレフィリングが忠実な思考（推論過程が回答の予測を忠実に説明していること）の生成を促すことによって性能が改善することを図示している．

[13, 14, 15, 16]．例えば，Sclar ら [13] は，Sally–Anne テストのような物語の文脈において登場人物の信念を追跡するためのパイプラインを設計した [17]．また，Wilf ら [14] は，質問に回答する前に，対象となる登場人物が何を知っているかに基づいて文脈をフィルタリングする手法を提案した．しかし，これらの手法はいずれも，世界の状態が変化する文脈における信念の推論に特化して設計されており[1)]，限られた問題設定にのみ適用可能である．一方で，様々な文脈で多様な種類の心的状態が ToM によって推論可能である [18]．したがって，特定の状況における信念のみに着目した手法では，LLM に実用的で人間らしい ToM を獲得させるには不十分である．

本研究では，様々な心的状態において LLM の ToM を向上させる新しい推論手法として，他人の靴

1）例：サリーがボールを箱に入れて部屋を出た後，アンがそのボールをかごに移した．サリーは部屋に戻ってきた時，ボールはどこにあると思うか？ [2].

（Shoes-of-Others, SoO）プレフィリングを提案する．SoO プレフィリングは，図 1 に示すように，他人の立場に立って推論することを促すために LLM の出力の冒頭を指定する（§2）．SoO プレフィリングは問題設定に関する仮定を最小限に抑えているため，より広範な設定に適用可能である．

本研究では，世界状態の変化を伴わない対話及び物語からなる文脈において，信念・意図・願望・感情・知識の 5 種類の心的状態にわたり，SoO プレフィリングが LLM の ToM を向上させ，他の手法を上回る性能を示すことを実証した（§3）．さらに分析の結果，SoO プレフィリングは，単に思考を長くすることに依存するのではなく（付録 C），より忠実な思考の生成を LLM に促すことで，性能改善に寄与している可能性が示唆された（§4）．

なお，本研究は EACL 2026 Findings に採択された研究 [19] に基づく．

2 他人の靴プレフィリング

SoO プレフィリング（図 1）は，出力の接頭辞として “Let’s put ourselves in {name}’s shoes.” を指定するだけの単純な手法である．{name} は質問によって心的状態の推論が求められている登場人物の名前を表す．本研究では，付録 B に詳述するルールベースの手法により，質問文から人物名を抽出した．その後，LLM はこの接頭辞に続く出力を生成する．

Wilf ら [14] と同様に，提案手法は，他者の視点に立って考えることを意味する視点取得（perspective taking） [20, 21] に着想を得て設計されている．視点取得は ToM 推論に不可欠であると考えられている [22]．Wilf ら [14] は，視点取得に基づき，登場人物が知らない情報を文脈から除外する手法を提案したが，この手法は登場人物の移動を伴う文脈にのみ適用可能である．これに対し，SoO プレフィリングは文脈に関する仮定を最小限に抑えているため，より広範な問題設定に適用可能である．

プレフィリングは，LLM に対して指示を与えるプロンプティングとは異なり，出力の先頭部分を指定してその続きを生成させるため，LLM の出力により強い制約をかけられる手法である．近年，LLM に有害な発話を生成させるためのプレフィリングが発見される [23] など，プレフィリングに関する研究への関心が高まっている．本研究の提案手法は，ToM 性能を改善することを目的とした，他者の視点に立つことを明示的に促すプレフィリングである．

3 実験

3.1 実験設定

データセット 本研究では，近年提案された 2 つのベンチマークである ToMATO [12] 及び ToMBench [10] を用いて ToM を評価した．これらのベンチマークは，いずれも世界状態の変化を伴わず，それぞれ対話的文脈及び物語的文脈において，幅広いカテゴリの心的状態にわたる ToM を評価するものである．ToMBench については，英語サブセットのみを用いた．評価指標としては，両ベンチマークが多肢選択式の質問応答として定式化されているため，正解率（accuracy）を用いた．各質問の選択肢数を 4 つに設定していることから，ランダムベースラインは 25% となる．ToMATO 及び ToMBench のデータサイズは，それぞれ 5.4k 及び 2.4k である．

手法 実験では，以下の 5 つの手法を比較した．(1) Vanilla：Zero-shot プロンプティング．(2) (Zero-shot) CoT プロンプティング [24]：入力末尾に “# Answer\n Let’s think step-by-step.” を付加する．(3) SoO プロンプティング：入力末尾に “Let’s put ourselves in {name}’s shoes.” を付加する．(4) CoT プレフィリング：出力の先頭に “Let’s think step-by-step.” を付与する．(5) SoO プレフィリング：提案手法．

モデル 3 つのオープンウェイト LLM（Mistral-7B-Instruct-v0.3 [25], Llama-3-8B-Instruct, 及び Llama-3-70B-Instruct [26]）と，プロプライエタリな LLM（GPT-4o mini）を使用した．

3.2 結果

一次の心の理論 一次の心の理論（first-order ToM）とは，一次の心的状態に関する推論を指す．例えば，“A thinks/will/wants/feels/knows X” は，それぞれ一次の信念/意図/願望/感情/知識に対応する．表 1 に示すように，SoO プレフィリングは，一部の例外を除き，対話入力（ToMATO）及び物語入力（ToMBench）の双方において，5 種類の心的状態で一貫して有効であった．一方で，プロンプティング手法は常に有効であるとは限らず，Vanilla と比較してスコアを低下させる傾向が見られた．

二次の心の理論 二次の心の理論（second-order ToM）とは，二次の心的状態に関する推論を指す．例えば，“A thinks that B thinks/will/wants/feels/knows Y” は，それぞれ信念/意図/願望/感情/知識に関する

表 1 ToMATO 及び ToMBench における一次の心の理論の性能（%）．3 回の実行における正解率の平均を報告する．B：信念，I：意図，D：願望，E：感情，K：知識．各モデルについて，最も高いスコアは **太字** で，Vanilla と比較して低下したスコアは 赤字 で示す．OpenAI のモデルはプレフィリングに非対応のため，プロンプティングのみを評価した．

		ToMATO						ToMBench					
Model	Method	B	I	D	E	K	Avg.	B	I	D	E	K	Avg.
GPT-4o mini	Vanilla	76.2	79.9	82.4	76.8	73.3	77.7	61.7	72.1	60.6	71.5	35.1	60.2
	CoT Prompting	48.1	46.4	51.7	62.6	43.8	50.5	44.9	30.5	34.4	46.6	28.4	36.9
	SoO Prompting	75.7	79.7	83.7	75.6	72.3	77.4	64.4	60.0	51.9	71.0	34.3	56.3
Mistral 7B	Vanilla	62.0	68.0	74.5	60.7	62.4	65.5	50.5	56.3	**51.0**	61.6	27.1	49.3
	CoT Prompting	62.7	**70.7**	74.5	62.4	**65.5**	67.2	53.7	58.5	50.2	60.5	30.6	50.7
	SoO Prompting	63.9	68.9	**76.5**	62.7	63.3	67.1	53.3	57.9	50.4	61.6	29.1	50.5
	CoT Prefilling	61.5	67.6	72.7	61.8	62.2	65.1	53.0	55.3	41.5	59.9	30.8	48.1
	SoO Prefilling	**64.9**	70.0	75.6	**63.0**	64.5	**67.6**	**56.2**	**58.9**	47.3	**63.7**	**34.5**	**52.1**
Llama3 8B	Vanilla	54.2	56.1	60.2	57.0	47.1	54.9	48.7	56.0	49.2	61.2	31.1	49.2
	CoT Prompting	26.0	26.2	22.0	28.9	24.7	25.6	46.1	41.9	36.9	51.8	28.0	40.9
	SoO Prompting	51.6	57.9	51.1	55.6	41.6	51.6	47.9	47.1	45.0	57.6	32.2	46.0
	CoT Prefilling	64.1	65.3	71.0	60.8	58.9	64.0	55.3	65.2	51.9	63.4	37.2	54.6
	SoO Prefilling	**67.2**	**69.2**	**73.4**	**65.7**	**62.0**	**67.5**	**61.1**	**70.8**	**59.0**	**66.6**	**38.3**	**59.2**
Llama3 70B	Vanilla	81.7	85.3	85.9	80.5	73.5	81.4	73.6	79.8	58.5	71.9	45.9	66.0
	CoT Prompting	80.5	85.2	86.7	81.3	74.1	81.6	68.2	78.8	54.0	69.6	50.6	64.2
	SoO Prompting	81.9	86.2	**87.6**	82.2	75.6	82.7	73.3	80.7	**59.4**	72.8	51.9	67.6
	CoT Prefilling	79.9	83.9	84.2	78.6	73.4	80.0	71.1	69.1	51.5	65.8	**49.0**	61.3
	SoO Prefilling	**82.2**	**86.9**	87.4	**82.4**	**76.7**	**83.1**	**80.5**	**80.8**	57.9	**73.0**	47.9	**68.0**

表 2 ToMATO における二次の心の理論の性能（%）．3 回の実行における正解率の平均を報告する．

		True Belief						False Belief					
Model	Method	B	I	D	E	K	Avg.	B	I	D	E	K	Avg.
GPT-4o mini	Vanilla	69.1	70.7	77.2	71.7	73.7	72.5	60.1	47.8	71.9	71.7	58.6	62.0
	CoT Prompting	35.2	46.3	45.4	51.9	36.5	43.0	32.9	35.2	47.5	52.0	31.5	39.8
	SoO Prompting	69.2	70.5	71.7	74.9	72.9	71.8	62.0	47.5	66.5	70.9	56.8	60.7
Mistral 7B	Vanilla	57.3	61.9	61.4	61.6	62.4	60.9	42.9	39.9	48.9	48.6	50.8	46.2
	CoT Prompting	55.2	**64.0**	**64.2**	63.1	61.9	**61.7**	41.9	41.5	**57.0**	48.8	53.1	48.5
	SoO Prompting	**58.9**	62.5	62.1	60.8	**64.3**	**61.7**	44.2	42.9	53.2	47.0	49.4	47.3
	CoT Prefilling	54.8	62.9	62.5	63.5	62.1	61.2	44.4	**43.4**	55.7	50.9	52.3	49.4
	SoO Prefilling	52.0	62.4	61.7	**64.1**	63.8	60.8	**47.5**	43.2	50.4	**54.6**	**56.6**	**50.5**
Llama3 8B	Vanilla	39.8	45.6	46.3	46.9	40.4	43.8	34.5	29.5	35.4	37.0	27.8	32.8
	CoT Prompting	25.6	23.9	25.0	27.3	23.4	25.1	25.6	21.6	29.3	27.0	22.8	25.3
	SoO Prompting	36.1	47.2	44.9	47.5	35.0	42.1	32.4	30.6	39.7	41.2	27.6	34.3
	CoT Prefilling	58.6	60.6	61.2	58.0	60.8	59.8	**49.5**	**46.2**	53.2	54.3	**52.9**	**51.2**
	SoO Prefilling	**60.1**	**61.4**	**64.4**	**59.9**	**61.8**	**61.5**	48.1	43.4	**55.1**	**56.4**	50.2	50.6
Llama3 70B	Vanilla	73.7	76.1	78.7	75.8	70.2	74.9	60.6	57.1	67.3	70.1	58.4	62.7
	CoT Prompting	77.0	76.5	78.9	79.7	75.7	77.6	**63.7**	**62.3**	68.3	74.8	63.2	66.5
	SoO Prompting	76.4	**79.4**	81.1	77.9	**77.2**	78.4	62.7	61.2	69.2	72.2	**64.2**	65.9
	CoT Prefilling	75.7	76.0	79.1	78.4	72.1	76.3	60.2	58.2	**71.7**	77.2	62.1	65.9
	SoO Prefilling	**79.5**	77.8	**81.2**	**81.5**	76.3	**79.3**	62.6	61.2	70.9	**78.0**	62.1	**67.0**

表 3 Llama-3-8B-Instruct を用いた Ablation study.

Method	Prefix	ToMATO	ToMBench
Ours	Let's put ourselves in {name}'s shoes.	62.8	63.2
- name	Let's put ourselves in others' shoes.	29.4	43.0
- name	Let's put ourselves in shoes of others.	16.2	28.8

二次の信念に対応する．また，“A thinks that B feels Y, while B feels X” という状況は，X = Y の場合には感情に関する真信念（true belief），X ≠ Y の場合には誤信念（false belief）に対応する．誤信念とは，事実とは異なる信念を指す．

表 2 に示すように，SoO プレフィリングは，Llama-3-70B において真信念と誤信念ともに平均値（Avg.）で他の手法を上回る性能を示した．プロプライエタリ及びオープンウェイトモデルのいずれにおいても，プロンプティング手法は Vanilla と比較してスコアを低下させる場合が多かったのに対し，SoO プレフィリングは一貫してスコアを向上させた．特に，SoO プレフィリングは，真信念よりも誤信念の理解において高い有効性を示した．

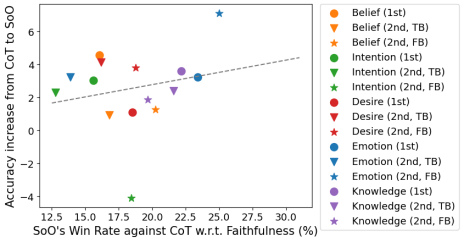
Ablation Study SoO プレフィリングにおいて登場人物名を含めることの効果を検証するため，Ablation Study を行った．表 3 に示すように，人物名を含めることは，2 つのベンチマークの双方において一貫して有効であることが確認された．ToM 推論における思考過程を適切に誘導するためには，接頭辞において人物名を明示的に指定することが重要であると考えられる．

4 分析

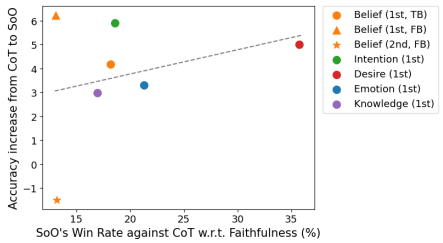
なぜ SoO プレフィリングは CoT プレフィリングを上回るのか？ Chain-of-Thought はしばしば非忠実性（unfaithfulness）の問題を抱えている．すなわち，生成された思考が LLM の最終的な予測を正確に説明していない場合があることが報告されている [27, 28]．SoO プレフィリングが §3 において他の手法を上回る性能を示した理由として，この非忠実性を緩和したことが寄与していると仮説を立てた．

この仮説を検証するため，まず GPT-4o mini を用いた LLM-as-a-judge [29] により，CoT 及び SoO プレフィリングの思考の忠実性を対比較した．対比較のために，Zheng ら [29] の対比較用プロンプトを拡張した．CoT と SoO の公平な比較のため，LLM-as-a-judge では，Llama-3-8B が CoT 及び SoO プレフィリングの双方で正しい予測を行った例のみを使用した．そして，正解率の向上と忠実性に関する勝率の間の相関を分析した．

図 2 に示すように，2 つのベンチマークの双方において，忠実性に関する勝率は正解率の向上と正の相関を示している．この傾向は他のモデルにおいても観察された．これは，SoO プレフィリングが思考の非忠実性を緩和し，その結果として ToM 性能を向上させていることを示唆している．LLM の出力に対する介入，思考の忠実性，及び ToM の関係を分析するこのような試みは，LLM における ToM 研究の中で比較的未開拓である．



(a) ToMATO



(b) ToMBench

図 2 Llama-3-8B-Instruct における正解率と忠実性の相関．両ベンチマークで正の相関を示す．

5 議論

本研究では，視点取得を明示的に促すことで LLM の ToM 性能が改善するという知見が得られた．これは，自閉スペクトラム症（ASD）の青年が，視点取得を促されない場合には定型発達者よりも社会的認知の成績が劣る一方で，明示的に促されることで両者が同等の成績を示すという ASD 研究の知見 [30] と一致している．また，回答する前に思考する（explicit ToM）ことで Zero-shot（implicit ToM）よりも性能が改善する点についても ASD 研究の知見 [31] と一致する．人間と LLM の比較を通して ToM の理解をさらに深めることは，今後の課題である．

6 結論

本研究では，適用可能な範囲を損なうことなく LLM の ToM を改善する手法として，SoO プレフィリングを提案した．2 つのベンチマークを用いた実験により，本手法が多くの場合において，5 種類の心的状態における一次及び二次，特に誤信念の理解において ToM 性能を向上させることを示した．さらに分析から，本手法が思考の非忠実性を緩和することで ToM 性能を改善していることが示唆された．

参考文献

[1] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? **Behavioral and Brain Sciences**, Vol. 1, No. 4, p. 515–526, 1978.

[2] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a “theory of mind”? **Cognition**, Vol. 21, No. 1, pp. 37–46, 1985.

[3] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. Gmail smart compose: Real-time assisted writing. In **KDD**, 2019.

[4] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In **WWW**, 2021.

[5] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In **ACL**, 2019.

[6] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In **EMNLP**, 2022.

[7] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[8] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In **EACL**, 2024.

[9] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In **EMNLP**, 2023.

[10] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In **ACL**, 2024.

[11] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In **ACL**, 2024.

[12] Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In **AAAI**, 2025.

[13] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In **ACL**, 2023.

[14] Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In **ACL**, 2024.

[15] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In **Findings of ACL**, 2024.

[16] Sneheel Sarangi, Maha Elgarf, and Hanan Salam. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In **COLING**, 2025.

[17] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In **EMNLP**, pp. 5872–5877, 2019.

[18] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. In **Findings of EMNLP**, 2023.

[19] Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Yoshihiro Yamazaki, Keita Suzuki, Hiroaki Sugiyama, and Kuniko Saito. Let's put ourselves in sally's shoes: Shoes-of-others prefilling improves theory of mind in large language models. In **Findings of EACL**, 2026.

[20] Mark H Davis, Laura Conklin, Amy Smith, and Carol Luce. Effect of perspective taking on the cognitive representation of persons: a merging of self and other. **Journal of personality and social psychology**, Vol. 70, No. 4, p. 713, 1996.

[21] Perrine Ruby and Jean Decety. How would you feel versus how do you think she would feel? a neuroimaging study of perspective-taking with social emotions. **Journal of Cognitive Neuroscience**, Vol. 16, No. 6, pp. 988–999, 07 2004.

[22] Chris D Frith and Uta Frith. The neural basis of mentalizing. **Neuron**, Vol. 50, No. 4, pp. 531–534, 2006.

[23] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In **ICLR**, 2025.

[24] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In **NeurIPS**, 2022.

[25] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[26] 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, 2024.

[27] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In **IJCNLP-AACL**, 2023.

[28] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **NeurIPS**, 2023.

[30] Björn Callenmark, Lars Kjellin, Louise Rönnqvist, and Sven Bölte. Explicit versus implicit social cognition testing in autism spectrum disorder. **Autism: The International Journal of Research & Practice**, Vol. 18, No. 6, 2014.

[31] Tobias Schuwerk, Maria Vuori, and Beate Sodian. Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. **Autism: The International Journal of Research and Practice**, Vol. 19, No. 4, pp. 459–468, 2015.

[32] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

A　実験設定

実験で用いたハイパーパラメータを表 4 に示す．

表 4　ハイパーパラメータ

ハイパーパラメータ	値
do_sample	True
top_p	0.9
temperature	0.6
max_new_tokens	1024

実験で用いたオープンウェイト LLM は以下の通り：Mistral-7B-Instruct-v0.3[2)], Llama-3-8B-Instruct[3)], Llama-3-70B-Instruct[4)] [26]．計算コストの低減のために，bitsandbytes[5)] による 4bit 量子化を用いた．プロプライエタリな LLM では，gpt-4o-mini-2024-07-18 を用いた．

データセットについて，ToMBench は 2 つの選択肢を持つ質問が含まれるが，単純のためにこれらはのぞいた．ToMBench について，英語のサブセットのみを用いた．

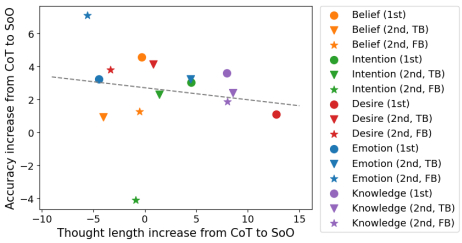
B　名前の抽出

SoO プレフィリングでは，質問によって心的状態の推論が求められている登場人物の名前を用いる．ToMATO では，質問がルールベースで生成されているため，登場人物名を決定的に特定することが可能である．一方，ToMBench では質問が手作業で作成されているため，本研究ではルールベースの手法を用いて質問文から人物名を抽出した．具体的には，spaCy[6)] を用いた品詞タグ付け及び依存構造解析により，対象となる登場人物名を特定した．この手法で人物名を特定できなかった質問は除外した．本研究ではルールベースの名前抽出手法を採用したが，In-Context Learning などの LLM を用いた手法も有用である可能性がある．LLM を用いた名前の抽出手法を検証することは，今後の課題である．

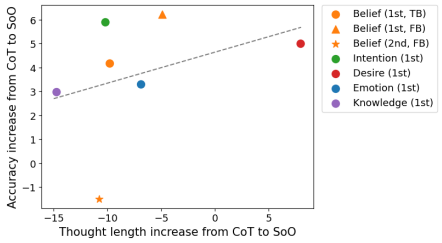
2)　https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
3)　https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
4)　https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
5)　https://github.com/bitsandbytes-foundation/bitsandbytes
6)　https://spacy.io/

C　分析

SoO プレフィリングは推論時の計算量をスケールすることで心の理論を改善しているか？　この問いに答えるため，CoT プレフィリングと比べた SoO プレフィリングの正解率の向上と思考の長さの増加についても相関分析を行った．図 3 に示すように，思考の長さの増加は必ずしも正解率の向上と正の相関を示さないことが分かる．この結果は，SoO プレフィリングが，思考を単に長くすること，すなわち推論時の計算量をスケールさせること [32] のみに依存して ToM を向上させているわけではないことを示唆している．



(a) ToMATO



(b) ToMBench

図 3　Llama-3-8B-Instruct における正解率と思考の長さの相関分析．両者の相関は必ずしも正ではない．