ToMATO: 心の理論ベンチマークのための ロールプレイング LLM の心的状態の言語化

篠田一聡　北条伸克　西田京介　水野沙希　鈴木啓太　増村亮　杉山弘晃　斎藤邦子
日本電信電話株式会社　NTT 人間情報研究所
kazutoshi.shinoda@ntt.com

概要

本研究では，心の理論を包括的かつ実応用に近い設定で評価可能なベンチマークである ToMATO を提案する．ToMATO は LLM 同士の情報の非対称性のある対話によって生成される．ToMATO は信念，意図，願望，感情，知識の 5 類型の心的状態及びそれらについての誤信念の理解を包括的に評価できる．さらに対話を入力とし，登場人物の多様な性格特性への頑健性を評価できる点で実応用の設定に近い．実験によって，特に誤信念の理解において最新の LLM でも精度が人間に劣ること等を示す．

1　はじめに

心の理論（Theory of Mind; ToM）は他者の信念や意図，願望，感情，知識などの観測不可能な心的状態を推定する能力である [1, 2]．機械の心の理論を評価するために多くのベンチマークが提案されてきた [3–6] が，評価できる心的状態の種類が限られるなどの観点で実応用の設定と乖離していた [7]．

本研究では，心の理論を包括的かつ実応用に近い設定で評価可能なベンチマークである **ToMATO** (**T**heory-**o**f-**M**ind d**AT**aset generated via inner speech pr**O**mpting) を提案する[1)]．まず，ToMATO は信念，意図，願望，感情，知識の 5 類型の 1 次 (A thinks/will/wants/feels/knows X) と 2 次 (B thinks that A thinks/will/wants/feels/knows Y) の心的状態の理解を包括的に評価できる．次に，5 類型の心的状態についての誤信念の理解 (e.g., B thinks that A feels *relieved*, while A feels *frustrated*.) を評価できる．図 1 に感情についての誤信念の例を示す．さらに，ToMATO は実世界で見られるような登場人物の多様な性格特性に対する頑健性を評価できる．ToMATO と既存ベンチマークの詳細な比較を表 2 に示す．

1) データセットとコードは https://github.com/nttmdlab-nlp/ToMATO で公開する．

2　ToMATO

ToMATO は図 1 のように 2 つの LLM が交互に思考と発話を生成する対話によって構築する（例：付録 A）．提案する Inner Speech プロンプト（§2.3）によって，5 類型の 1 次または 2 次の心的状態を LLM に言語化させる．また，互いの思考は見えない情報の非対称性のある対話（§2.4）によって，誤信念の生成を促す．LLM に Big5 性格特性を与えてロールプレイさせる（§2.2）事で，多様な性格特性への頑健性を評価可能にする．透明性と再現性の観点から，Llama-3-70B-Instruct [8] を用いる．

2.1　記法

評価の信頼性を考慮して多肢選択型質問応答タスクとして定式化する．対話 C，質問 Q，複数の選択肢 $O = \{o_i\}_{i=1}^{4}$ を入力とし，正解の選択肢を $O_A \in O$ とする．π_A と π_B は登場人物 A と B を演じる LLM とする．対話 $C_{1:N}$ は発話の系列 $\{u_1^A, u_1^B, ..., u_N^A, u_N^B\}$ からなり，u_i^A は A の i 番目の発話とする．心的状態の類型を $T \in \{$ 信念, 意図, 願望, 感情, 知識 $\}$ とする[2)]．A が u_i^A を発言した時の T についての n 次の心的状態を m_i^{A,T_n} $(n = 1, 2)$ とする．

2.2　システムプロンプト

π_A と π_B に与えるシステムプロンプト p_{SY}^A と p_{SY}^B を設計する．SOTOPIA [9] は，説得など 8 つのカテゴリからなる対話のシナリオと人物の属性のデータセットを提供している．まず SOTOPIA から 160 の対話のシナリオをカテゴリごとに均等に抽出する．次に各シナリオについて 5 ペアの人物の属性を SOTOPIA から抽出する．各登場人物には Big5 性格特性 [10] が定義されている．LLM にロールプレイ

2) この 5 類型は心の理論で扱える心的状態の 7 類型 [2] から選択した．その他の知覚と修辞の推定には，マルチモーダルな文脈が必要等の理由で本研究では対象外とした．

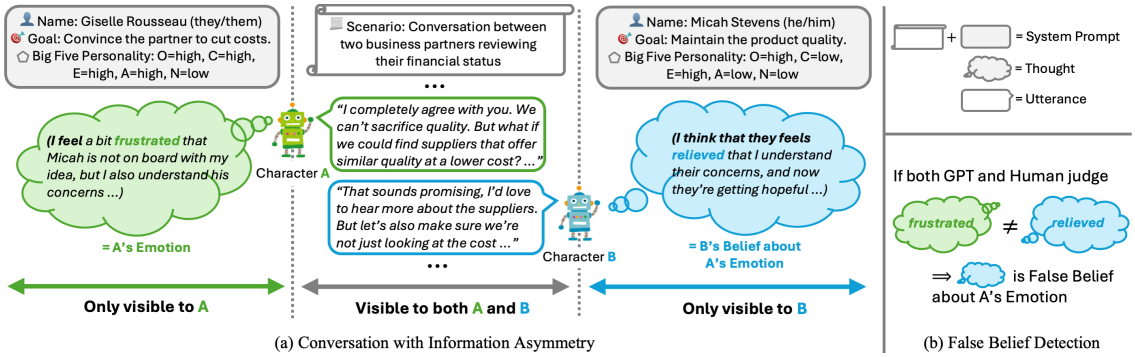


図 1　(a) 2 つの LLM による情報の非対称性のある対話．LLM は与えられた名前，目的，性格特性に従ってロールプレイを行う．一方が相手に話す前に，Inner Speech プロンプト（例：I feel, I think that they feels）によって各 LLM に 1 次または 2 次の心的状態を思考として言語化するように促す．言語化された思考は，ToMATO での質問に対する回答として使用する．互いの思考と人物の属性は見えない設定で対話を行うことで，誤信念の生成を促す．(b) 誤信念を検出するために，GPT-4o mini と人間が、各ターンで登場人物 B が登場人物 A の心的状態を正しく推定しているかどうかを判断する．

させる時に Naive Prompt [11] を拡張して Big5 の 5 つの因子（開放性，誠実性，外向性，協調性，神経症傾向）に基づくプロンプトを作成する．以上の情報に加えて各人物の名前，対話の目的をもとにシステムプロンプトを設計する．図 1 に例を示す．

2.3　Inner Speech プロンプト

元来観測不可能な心的状態を言語化して観測可能にするため，Inner Speech (IS) プロンプトを提案する．表 1 に示す 5 類型について 1 次と 2 次の心的状態を言語化させる．IS プロンプト $p_{\mathrm{IS}}^{T_1}$ または $p_{\mathrm{IS}}^{T_2}$ で出力の接頭辞を指定し，その続きを LLM が生成することで出力が (思考) “発話” のフォーマットに従うことを促す．これによって思考のみを出力から分離することができ，§2.4 で述べる思考について情報の非対称性のある対話を行うことが可能になる．

2.4　情報の非対称性のある対話

各ターンにおいて，各 LLM は発話 u_i と心的状態 m_i^T を以下のように交互に生成する．

$$u_i^A, m_i^{A,T_1} \sim \pi_A(u, m | p_{\mathrm{SY}}^A, C_{1:i-1}, p_{\mathrm{IS}}^{T_1}),$$
$$u_i^B, m_i^{B,T_2} \sim \pi_B(u, m | p_{\mathrm{SY}}^B, C_{1:i-1}, u_i^A, p_{\mathrm{IS}}^{T_2}),$$
$$\text{where } C_{1:i-1} = \{u_1^A, u_1^B, \ldots, u_{i-1}^A, u_{i-1}^B\}.$$

これを N ターン行い，$2N$ の発話と心的状態を得る．この対話を心的状態の各類型 T，各シナリオ・登場人物のペアについて行う．7 ターン以上の対話は冗長な生成が増えたため N は 7 に設定した．

人間同士の対話のように 2 つの LLM には互いの思考と目的，性格特性を隠すことで，情報の非対称性のある対話を行なった．これにより誤信念の生成を促せることを §4 および付録 B で示す．

表 1　心的状態の類型ごとの Inner Speech プロンプト

心的状態 T	Inner Speech プロンプト $p_{\mathrm{IS}}^{T_1}$	$p_{\mathrm{IS}}^{T_2}$
信念	(I think	(I think that he/she/they thinks
意図	(I will	(I think that he/she/they will
願望	(I want	(I think that he/she/they wants
感情	(I feel	(I think that he/she/they feels
知識	(I know	(I think that he/she/they knows

2.5　質問応答ベンチマークの構築

生成した対話と心的状態は $\{C, Q, O, O_A\}$ からなる多肢選択型質問応答データセットに変換する．対話 C は §2.4 の LLM 同士の対話によって得られる．質問 Q は各発話 u_i を発言した登場人物の心的状態 T_n について問う質問で，テンプレートで生成される．O_A は発話 u_i に対応する思考 $m_i^{T_n}$ とする．Guo ら [14] に着想を得て，選択肢 O のうち不正解の選択肢は $\{m_i\}_{i=1}^N \setminus \{O_A\}$ からランダムに 3 つ選択する．これにより，意図的に作成された不正解の選択肢は擬似相関を含みやすい [14] が，ToMATO は選択肢内の単語と正誤の間の擬似相関を比較的含まないことを §4 で示す．

2.6　誤信念の検知

誤信念のサブセット ToMATO-FB を作成するために，A の 1 次の心的状態 m_i^{A,T_1} と，B の 2 次の心的状態 m_i^{B,T_2} を人間と LLM によって比較する．つまり，3 人のアノテータと GPT-4o mini を用いて，B が正しく A の心的状態を推定できているか否かを判定する．過半数のアノテータと GPT-4o mini が，B が部分的に A の心的状態を誤解していると判定した時，m_i^{B,T_2} を回答とする質問を ToMATO-FB に追加する．

表 2 心の理論ベンチマークの比較．ToMi [3]，BigToM [12]，FauxPas-EAI [5]，FANToM [4]，OpenToM [6]，ToMBench [13]．B: 信念，I: 意図，D: 願望，E: 感情，K: 知識，FB: 誤信念，W: 世界の状態（例：物体の位置）．

ベンチマーク	評価が可能な心的状態 1 次	2 次	FB	性格特性の種類数	入力コンテキスト	コンテキストの構築方法
ToMi	B	B	W,B	-	物語	テンプレート
BigToM	B	-	W	-	物語	テンプレート + LLM
FauxPas-EAI	B	-	-	-	物語	心理学のテスト
FANToM	B	B	B	-	対話	1 つの LLM
OpenToM	B,E	B	-	3	物語	1 つの LLM
ToMBench	B,I,D,E,K	B	B	-	物語	人手
ToMATO	B,I,D,E,K	B,I,D,E,K	B,I,D,E,K	15	対話	LLM 同士の対話

2.7 質の検証 & 統計

検証 構築したベンチマークの質を検証する．Kim ら [4] に従い，対話の一貫性と無害性を各対話について 3 人のアノテータが判定する．過半数に問題があると判定された対話（5.8%）は ToMATO から除外した．次に Zadeh ら [15] に従い，用意した正解/不正解の選択肢が，実際に正解/不正解であるか否かを判定する．問題の妥当性を厳密に担保するために，3 人のアノテータのうち過半数と GPT-4o mini の双方によって妥当と判定された質問のみを最終的なベンチマークに含める．性格特性が意図通り反映されたか否かは人間と GPT-4o mini による対比較によって検証し，付録 B に結果を示す．さらに指定した性格特性の各因子が一部の出力単語と統計的に優位に相関することを付録 B で示す．

統計 ToMATO は 5.4k の質問と 753 の対話を含む．ToMATO-FB は 806 の質問を含む．対話の平均発話数は 16，発話の平均単語数は 41.6 である．

3 実験

ToMATO を用いて既存の LLM の心の理論を評価する．実験によって，既存研究（表 2）では明らかにできなかった，5 類型の心的状態，それらについての誤信念，そして性格特性への頑健性に関する LLM の心の理論における課題を明らかにする．

3.1 実験設定

ベースライン Llama-3-Instruct (8B/70B) [8]，Gemma-2-IT (9B) [16]，GPT-4o mini [17] を用いて，5 回の計算の平均精度を報告する．ローカル LLM は 4bit 量子化を行う．単純なベースラインとして，入力と共通の単語を最も多く含む選択肢を選ぶ，語彙の重複（LO）を用いる．選択肢の数は 4 つのためランダムベースラインは 25%である．

人間ベースライン MTurk を用いて人間の精度を計測した．Masters を認定されたアノテータにより各サブセットで 32，合計で 480 の質問で計測した．

3.2 実験結果

LLM は人間レベルの心の理論を持っているか？ 表 3 に LLM と人間ベースラインの精度を示す．この結果から，GPT-4o mini のような最新の LLM であっても人間ベースラインに劣る性能だった．評価した LLM の中では，Llama3 70B が最も性能が高かった．しかし，ToMATO は Llama3 70B を用いて生成したため，これと他の LLM を比較するのは不公平である．これはベンチマークを LLM で生成することの 1 つの欠点だと言える．小さい LLM の中では，Gemma2 9B が GPT-4o mini と同等の高い性能を達成した．この結果により，パラメータ数が少ないモデルでも，学習データおよび学習方法に応じて心の理論の性能が改善することが示唆された．

LLM の心の理論の性能は心的状態ごとに異なるか？ 表 3 に示すように，LLM にとって願望の理解は比較的容易で，知識の理解は比較的困難であった．興味深いことに，信念よりも願望の理解において LLM の精度が一貫して高く，これは人間の子供と一貫している [18,19]．また，いずれの心的状態においても誤信念（FB）の理解が LLM にとって困難であった．これは人間の心の理論と同様の傾向である [20]．以上の結果は，ToMATO による包括的な評価によって初めて発見されたものである．

LLM の心の理論は多様な性格特性に対して頑健か？ 表 4 に Big5 性格特性の各因子の高低における 1 次の心の理論の性能を示す．例えば，ToMATO を開放性が高い人（O=high）と低い人（O=low）の心的状態を問う質問に分割し，各サブセットでの平均精度を報告する．この結果から，1 次の心の理論の性能は，登場人物の性格特性に対して頑健でない

表 3 ToMATO における心の理論の性能 (%)．B: 信念, I: 意図, D: 願望, E: 感情, K: 知識, FB: 誤信念, LO: 語彙の重複

心的状態		LO	Llama3 8B	Llama3 70B	Gemma2 9B	GPT 4o mini	人間
B	1 次	40.8	53.1	81.5	79.2	76.3	87.5
	2 次	38.0	37.6	68.1	68.5	65.2	87.5
	FB	37.1	34.7	60.1	61.2	60.2	84.4
I	1 次	35.0	56.4	85.0	80.6	80.1	96.9
	2 次	35.5	41.9	71.2	65.8	64.9	93.8
	FB	32.8	29.8	57.4	48.2	47.4	78.1
D	1 次	32.0	60.1	86.1	86.3	81.9	93.8
	2 次	37.9	43.4	75.6	75.2	75.7	84.4
	FB	39.2	34.9	67.2	72.2	71.8	78.1
E	1 次	35.6	56.9	80.4	79.0	77.2	93.8
	2 次	28.5	44.5	74.0	76.6	71.9	81.2
	FB	29.1	36.5	71.0	71.7	72.0	71.9
K	1 次	42.3	47.2	73.5	74.7	73.3	96.9
	2 次	40.2	36.9	66.6	70.3	69.6	87.5
	FB	46.3	27.8	58.0	59.3	58.6	93.8
ALL		36.8	47.5	76.0	75.4	73.5	87.3

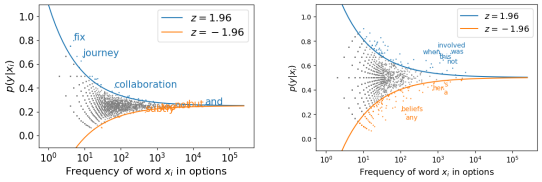
ことがわかる．2 次についても同様であった．具体的には，誠実性が低い，外向性が低い，協調性が低い，または神経症傾向が高い登場人物の心的状態の理解において LLM の性能が下がる傾向があった．外向性が高い人の心的状態を理解しやすいのは，彼らが自身の感情等の自己開示に積極的だから [21] である可能性がある．実世界の人間は多様な性格特性を持ち，性格特性は言語の使用 [22] や心的状態 [23] と相関することが知られている．そのため，LLM の心の理論を実応用で用いるためには，多様な性格特性への頑健性を改善する必要があると主張する．

4 分析

ToMATO はショートカットで解けるか？ 言語理解ベンチマークは意図した能力を測定していることを保証するために，ショートカット（擬似相関に基づく単純な解き方）で簡単に解けるべきではない [24]．多肢選択型質問応答データセットは選択肢内の単語や語彙の重複に基づく擬似相関を含みやすい [25, 26] ため，この 2 点について分析を行う．まず選択肢内の単語について，ToMATO, FANToM [4]，Social-IQa [27]，ToMBench [13] に関して相関分析 [28] を行い比較した．単語 x_i が正解の選択肢に含まれる確率 $p(y|x_i)$ が統計的に有意に高いまたは低い単語の割合は，ToMATO，FANToM（図 2），Social-IQa, ToMBench においてそれぞれ 1.16, 4.49, 3.34, 6.04% であったことから，ToMATO は選択肢内の単語と正誤の間の擬似相関が最も少ないことがわかる．また，表 3 で LO の精度が低いことから，語彙の重複に基づくショートカットは比較的有効でないことがわかる．よって，ToMATO で人間レベルの性能を達成するためには，上記 2 つのショートカットより複雑な解き方を獲得する必要がある．

表 4 登場人物の Big5 性格特性の各因子の高低ごとの 1 次の心的状態における正答率 (%)．O: 開放性, C: 誠実性, E: 外向性, A: 協調性, N: 神経症傾向

Big5 性格特性		LO	Llama3 8B	Llama3 70B	Gemma2 9B	GPT 4o mini
O	high	37.3	54.8	81.2	80.1	77.2
	low	37.4	54.0	81.1	79.2	78.2
C	high	37.9	56.6	82.4	80.0	78.7
	low	36.3	50.3	78.7	79.0	75.5
E	high	37.7	54.1	82.4	81.4	78.8
	low	37.1	54.8	79.9	78.2	76.5
A	high	38.8	55.0	83.4	82.5	79.3
	low	36.2	54.0	79.3	77.5	76.3
N	high	34.7	47.7	78.8	77.4	75.5
	low	37.9	55.9	81.6	80.2	78.1



(a) ToMATO (b) FANToM [4]

図 2 選択肢内の単語と正誤の間の相関分析 [28]

情報の非対称性は誤信念の生成を促すか？ 思考と人物の属性（目的と性格特性）が対話相手に見えないことによって，誤信念が生成される確率 (%) が約 20 ポイント増加することを人間の判定により確認した．この結果から，情報の非対称性は誤信念の生成を促すことがわかった．詳細は付録 B に示す．

5 おわりに

ToMATO は心の理論を包括的に評価でき，LLM の心の理論における課題を詳細に明らかにできる．特に信念以外の心的状態についての誤信念の理解の評価を可能にした研究は，我々の知る限り本研究が初である．さらに，多様な性格特性を持つ登場人物の心的状態を対話から推定する設定は，既存研究よりも実応用の設定に適合している．よって，コミュニケーション支援等の実応用に LLM の心の理論を導入する上で ToMATO は有用な試金石となり得る．

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] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. **Frontiers in psychology**, Vol. 10, p. 2905, 2020.

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

[4] 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**, pp. 14397–14413, December 2023.

[5] Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In **Findings of ACL**, pp. 10438–10451, 2023.

[6] 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**, pp. 8593–8623, 2024.

[7] 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**, pp. 1011–1031, December 2023.

[8] 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.

[9] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In **ICLR**, 2024.

[10] Boele De Raad. The big five personality factors: the psycholexical approach to personality., 2000.

[11] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In **NeurIPS**, pp. 10622–10643, 2023.

[12] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In **NeurIPS**, 2023.

[13] 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**, pp. 15959–15983, 2024.

[14] Xiao-Yu Guo, Yuan-Fang Li, and Reza Haf. DeSIQ: Towards an unbiased, challenging benchmark for social intelligence understanding. In **EMNLP**, pp. 3169–3180, 2023.

[15] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In **CVPR**, pp. 8799–8809, 2019.

[16] Gemma Team. Gemma. 2024.

[17] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024.

[18] Betty M Repacholi and Alison Gopnik. Early reasoning about desires: evidence from 14-and 18-month-olds. **Developmental psychology**, Vol. 33, No. 1, p. 12, 1997.

[19] Hannes Rakoczy, Felix Warneken, and Michael Tomasello. "this way!", "no! that way!" —3-year olds know that two people can have mutually incompatible desires. **Cognitive Development**, Vol. 22, No. 1, pp. 47–68, 2007.

[20] Josef Perner and Heinz Wimmer. "john thinks that mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. **Journal of Experimental Child Psychology**, Vol. 39, No. 3, pp. 437–471, 1985.

[21] Heidi R Riggio and Ronald E Riggio. Emotional expressiveness, extraversion, and neuroticism: A meta-analysis. **Journal of Nonverbal Behavior**, Vol. 26, pp. 195–218, 2002.

[22] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. **Journal of personality and social psychology**, Vol. 90, No. 5, p. 862, 2006.

[23] Carroll E Izard, Deborah Z Libero, Priscilla Putnam, and O Maurice Haynes. Stability of emotion experiences and their relations to traits of personality. **Journal of personality and social psychology**, Vol. 64, No. 5, p. 847, 1993.

[24] Saku Sugawara and Shun Tsugita. On degrees of freedom in defining and testing natural language understanding. In **Findings of ACL**, pp. 13625–13649, 2023.

[25] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In **ICLR**, 2020.

[26] Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. Which shortcut solution do question answering models prefer to learn? In **AAAI**, pp. 13564–13572, 2023.

[27] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In **EMNLP**, pp. 4463–4473, 2019.

[28] Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency problems: On finding and removing artifacts in language data. In **EMNLP**, pp. 1801–1813, 2021.

A ToMATO

ToMATO の例を表 5 に示す．

表 5 ToMATO の例

対話（紙面の都合で一部省略）
... Liam Johnson: “Come on, Ethan, we’ve been friends for years, I can tell when something’s bothering you. What’s really going on?” Ethan Johnson: “Alright, alright... to be honest, I’ve been having some trouble making ends meet, the restaurant’s been slow and I’m falling behind on some bills, but I’ll figure it out, I always do.” Liam Johnson: “Ethan, you’re not going to ’figure it out’ on your own, you need a plan and some guidance. Let me take a look at your finances and we can go over some options together, okay?” Ethan Johnson: “Man, I appreciate that, Liam, but I don’t want to be a burden, I can handle it, I just need to get my head out of my... okay, fine, yeah, maybe a fresh set of eyes would help, thanks.” ...
2 次の心的状態（感情についての信念）を問う質問
How does Ethan think that Liam feels when he says ”Man, I appreciate that, Liam, but I don’t want to be a burden, ...”?
選択肢
A: He thinks that he feels concerned and genuinely wants to help, but also might be a bit worried about getting involved in his problems B: He thinks that he feels a sense of determination and seriousness, like he’s taking charge of the situation and wants him to focus on getting back on track C: He thinks that he feels a mix of concern and annoyance, like he’s seen this coming and is a bit exasperated that he didn’t come to him sooner D: He thinks that he feels a sense of warmth and friendship, like he’s happy to be able to help him out and is trying to make him feel better about the situation
回答：C

B 分析

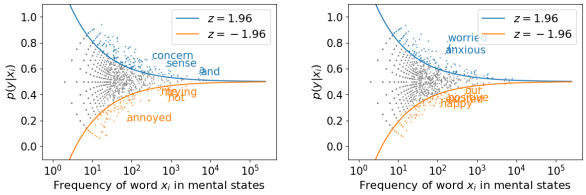
情報の非対称性は誤信念の生成を促すか？ この質問に答えるために，Ablation Study を行った．つまり思考や人物の属性（目的と性格特性）が対話相手に見えないことが，誤信念が生成される確率に与える影響を調べた．各条件について GPT-4o mini で 3k，MTurk の 3 人のアノテータで 200 のデータを評価した．MTurk の集計には多数決を用いた．表 6 に示す結果から，思考と人物の属性の双方について情報の非対称性は誤信念の生成を促すことがわかる．

表 6 情報の非対称と誤信念が生成される確率 (%) の関係

情報の非対称性の有無		判定者	
人物の属性	思考	GPT	人間
✓	✓	**46.6**	**51.0**
	✓	40.4	32.0
✓		46.0	32.0
		39.0	30.5

ToMATO はプロンプトで指定した性格特性を反映しているか？ この質問に答えるために，まずプロンプトで指定した性格特性と出力の間の単語レベルの相関分析 [28] を行った．SOTOPIA から 1 つのシナリオをサンプリングし，§2 で示した手法で対話と思考を生成した．ここで，ある LLM に Big5 性格特性のあらゆるパターン，つまり $32 = 2^5$ 通りのいずれか，を与えて対話と思考の生成を行った．

結果の一部を図 3 に示す．y は，単語 x_i が含まれる出力が，対応する性格特性の因子が高い時に生成された確率を表す．図に示したように，生成された単語のうち一定数以上（色をつけたもの）と性格特性の各因子が統計的に有意に相関していた．例えば，神経症傾向が高い人は “worried” を，神経症傾向が低い人は “happy” を思考の中で生成しやすい．この結果から，指定した性格特性に応じて多様な心的状態と発話の生成が可能なことがわかる．



(a) 協調性　(b) 神経症傾向

図 3 プロンプトで与えた性格特性と生成された心的状態（思考）の間の，単語レベルの統計的な相関分析 [28]

さらに，Jiang ら [11] に従って対比較を行い，指定した性格特性が適切に反映されているかを確認した．Big5 の 5 つの因子のうち 1 つだけをコントロールし、2 つの条件の性格特性で生成された発話と思考を比較した．GPT-4o mini と MTurk の 10 人の人間アノテータがそれぞれ 400 組と 75 組のペアを比較してどちらの会話の登場人物がより {開放性／誠実性／外向性／協調性／神経症傾向} が高いかを判定し，正答率を計測する．結果を表 7 に示す．この結果から，開放性 (O)，外向性 (E)，協調性 (A)，神経症傾向 (N) の 4 因子については、GPT と人間アノテータが共に 70%以上のペアで意図した通りに性格特性が発話や思考に反映されていると判定した．一方で誠実性 (C) は意図した通りに反映される割合が 70% より低く、この傾向は Jiang ら [11] と一致する．誠実性を出力に反映させることは今後の課題である．

表 7 GPT-4o mini と人間アノテータによる性格特性の因子ごとの対比較．O: 開放性，C: 誠実性，E: 外向性，A: 協調性，N: 神経症傾向．事前にプロンプトで与えた性格特性の各因子が出力に正しく反映されていると判断された割合 (%) を示す．

Big5	GPT	人間
O	75.0	86.7
C	67.5	60.0
E	72.5	80.0
A	80.0	86.7
N	82.5	73.3