Theory of mind のベンチマーク指標は対話能力と関係があるのか？ LLM における対話能力と Theory of Mind の相関分析

伊勢野晴久 [1,2]　大橋厚元 [1]　小川哲司 [3]　高道慎之介 [4]　東中竜一郎 [1,2]
[1] 名古屋大学大学院 情報学研究科　[2]NII LLMC [3] 早稲田大学 [4] 慶應義塾大学
{iseno.haruhisa.h4, ohashi.atsumoto.c0}@s.mail.nagoya-u.ac.jp
ogawa.tetsuji@waseda.jp, shinnosuke_takamichi@keio.jp
higashinaka@i.nagoya-u.ac.jp

概要

対話では，Theory of Mind（ToM）による相手の心的状態の推測が重要な役割を果たすと考えられる．大規模言語モデル（LLM）においても，ToM の向上が対話能力の改善をもたらすと期待されているが，ToM と対話能力の関係性についての定量的検証は不十分である．そこで本研究では，7 種類の高性能な LLM を対象に，3 つの ToM ベンチマーク（ToMBENCH，FANToM，Hi-ToM）と 6 つの対話タスクにおける性能を評価することで，ToM 性能と対話能力の相関関係を実験によって調査した．その実験の結果，ToM と対話能力の間に相関関係が確認された．その一方で，評価する ToM の側面によって相関の強さに違いが生じることも分かった．具体的には，会話形式で評価される ToM や直接信念を尋ねる質問形式で評価される ToM は対話能力とより高い相関を示すことが分かった．

1　はじめに

近年，大規模言語モデル（LLM）をベースとした対話システムは，多様な対話タスクにおいて目覚ましい性能向上を示している [1, 2]．より人間らしい高度な対話能力を実現するためには，単なる言語処理能力の向上だけでなく，相手の心的状態を理解し推論する能力，すなわち Theory of Mind（ToM）が不可欠だと考えられている [3, 4]．

LLM の ToM を向上させるため，ToM を評価するベンチマークが数多く提案されてきた [5, 6, 7, 8, 9, 10, 11]．これらのベンチマークでは，LLM に物語や対話履歴を提示し，登場人物の信念や意図などの心的状態に関する質問応答を行うことで ToM を評価する．こうしたベンチマークによって ToM を向上させることで，タスク達成に向けて相手とコミュニケーションを行う対話能力の向上が期待されている．しかし，こうした期待がある一方で，LLM の ToM ベンチマーク性能と対話能力の関係性は定量的に検証されておらず，LLM の ToM ベンチマークにおける性能改善が対話能力の改善につながるかは明らかではない．

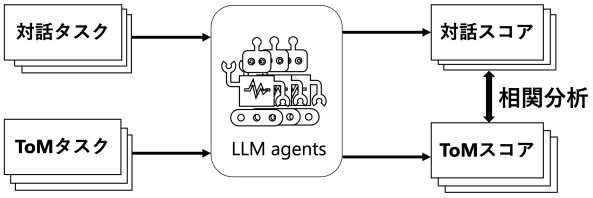

図 1　LLM における対話タスク性能と ToM ベンチマーク結果の相関を分析する実験フレームワーク

そこで，本研究では既存の ToM ベンチマークが対話において必要となる ToM をどの程度正確に捉えられているかを定量的に検証する．具体的には，7 種類の最先端 LLM を対象に，3 つの異なる ToM ベンチマーク（Hi-ToM [12]，FANToM [13]，ToMBENCH [14]）と 6 つの対話タスク（Taboo，Wordle，Drawing，Reference Game，Private & Shared，MutualFriends）での性能を評価し，ToM と対話能力の相関関係を調査する．

2　アプローチ

本研究では，LLM の ToM が実際の対話能力とどのような関係にあるかを明らかにするため，LLM の ToM と対話能力の相関分析を行う．図 1 に示すように，n 個の最先端の LLM に対して m 個の ToM ベンチマークと l 個の対話タスクを実行させ，LLM の ToM ベンチマーク正解率と対話タスクのスコア間の相関係数を算出する．

2.1 ToM ベンチマークの選定

ToM の全体的な性能を評価できることも重要だが，ToM の様々な観点についても性能が評価できると，対話能力と ToM の相関について包括的な評価ができると考えられる．そこで，本研究では，対話能力と特に関連すると考えられる観点として，以下の 3 つの観点に着目してベンチマークを選定する．

第一の観点はコンテキスト形式である．ToM ベンチマークには，Sally-Anne 課題 [15, 16] のように物語文から登場人物の信念推定を行うタスクと，会話文から登場人物の心的状態を推論するタスクがある．また，物語形式であっても，物語文の中に登場人物間の対話を含む設定と含まない設定が存在する．本研究では，これらのコンテキスト形式の違いが対話能力との相関に与える影響を分析するため，コンテキストが物語形式の場合と対話形式の場合，そして物語の中に対話を含む場合と含まない場合を網羅できるようにベンチマークを選定する．

第二の観点は質問形式である．ToM ベンチマークでは多様な質問形式があり，例えば「A は X がどこにあると思っているか」のような信念を直接問う形式と，「X がどこにあるか知っている人物はだれか」のような知識保有者の特定を求める形式がある．これらの質問では回答内容は異なるが，正しく解答するためには「ある人物は X がどこにあると思っているのか」という同じ ToM が必要となる．本研究では，必要な ToM は同じであっても，質問形式の違いが対話能力との相関に影響するかを調査するため，異なる質問形式を含むベンチマークを選定する．

第三の観点は推論次数である．1 次信念は「A は X だと思っている」という人物自身の信念であり，2 次信念は「B は，A が Y だと思っていると思っている」という他者の信念についての信念である．さらに，一部の ToM ベンチマークでは 3 次信念以上のより高次の推論能力も測定される．本研究では，これらの 1〜4 次信念それぞれに対する ToM と対話能力との相関を分析するため，複数の次数の信念に対する ToM 推論が含まれるベンチマークを選定する．

2.2 対話タスクの選定

本研究では，ToM と対話能力の相関を明らかにするため，タスク達成度を明確な評価指標で定量化でき，かつ ToM がタスク達成に重要となるタスク指向対話タスクを選定する．

表 1 本研究で扱う ToM ベンチマークに含まれる観点

	Hi-ToM	FANToM	ToMBENCH
コンテキスト形式	物語	対話	物語
質問形式	-	3 種類	-
推論次数	0-4	1-2	1-2

3 実験

ToM と対話能力の相関を調査するために実験を行った．まず，相関分析に用いる ToM ベンチマークと対話タスクを具体的に選定した．そして，7 つの最先端の LLM （GPT-4.1，Gemini 2.5-Flash，Claude 4-Sonnet，Grok-4，Llama 3.3-70B，Qwen 3-32B，Mistral-Small）に ToM ベンチマークの問題を回答させ，同じ LLM 同士の対話もしくは LLM とユーザシミュレータの対話によって対話タスクを実施した．その後，ToM と対話能力の相関を確認した．

3.1 ToM ベンチマーク

本研究では，表 1 に示す 3 つの ToM ベンチマーク （Hi-ToM，FANToM，ToMBENCH）を選定した．これらのベンチマークは 2.1 節で示したように，コンテキスト形式，質問形式，推論次数の違いを網羅的にカバーしており，各観点での相関分析を可能にする．これらの 3 つの ToM ベンチマークにおける LLM の正解率を LLM の ToM の指標として用いる．

コンテキスト形式の違いによる影響の分析には，物語形式の代表的なベンチマークとして ToMBENCH と Hi-ToM を，対話形式の代表的なベンチマークとして FANToM を活用する．また，Hi-ToM には物語中に登場人物間のコミュニケーションを含む設定（Tell）と含まない設定（No_Tell）の 2 種類の設定が存在する．これらの設定を比較することによって，物語形式のベンチマークにおけるコミュニケーション要素の有無による相関変化も分析する．

質問形式の違いによる影響の分析については，FANToM の 3 つのサブタスクを活用する．(1) BeliefQA：登場人物の信念状態を直接問う課題，(2) InfoAccessibilityQA (InfoQA)：特定の情報にアクセス可能な人物を列挙する課題，(3) AnswerabilityQA (AnsQA)：質問に正しく答えられる人物を列挙する課題の 3 つである．これらは同一の ToM 推論を異なる質問形式で問うため，質問形式の違いによる相関への影響を分析できる．

推論次数の違いによる影響の分析には，Hi-ToM の

1-4 次 ToM 推論タスク，また ToMBENCH と FANToM の 1 次・2 次信念推定タスクを利用する．これらの次数ごとの相関を比較することで，推論次数の違いによる相関への影響を分析する．

3.2 対話タスク

本研究では LLM の対話能力を評価するため，LLM の対話評価ベンチマークである Clembench [17] で用いられた 5 種類のテキストゲーム（Taboo, Wordle，Drawing，Reference Game，Private & Shared）を選定した．また，本研究では MutualFriends [18] タスクを 6 つ目の対話タスクとして選定した．これらのタスクは全て定量的な評価指標を持つタスクである．また，これらのタスクは相手がどの情報にアクセスできるかを考慮したり，自分の心的状態を積極的に相手に伝える必要があることから，タスク達成には ToM が重要と考えられる．

LLM 同士もしくは LLM とルールベースのユーザシミュレータの対話を通じて対話タスクを実施したのち，各対話タスクにおける LLM の性能を 0-100 点でスコアリングすることで，対話能力の指標として用いる．MutualFriends では対話の成功率のみを用いてスコアを算出し，その他のタスクでは Clembench で定義されたスコアリング方法を適用した．さらに，6 つのタスクの平均スコアを Average として算出し，LLM の全体的な対話能力を評価する指標として活用する．

3.3 実験手順

ToM ベンチマークの実行では，LLM にコンテキストとなる物語文もしくは対話文を提示し，その上で人物の心的状態に関する問題を選択問題式で回答させた．この際，Hi-ToM はデータセットに含まれる既存の問題回答用のプロンプトをそのまま利用し，ToMBENCH と FANToM については，本研究でプロンプトを新たに設計した（付録 A）．

対話タスクの実行では，同じ LLM 同士で対話を行わせた．Clembench に含まれる 5 つのタスク（Taboo, Wordle, Drawing, Reference Game, Private Shared）を行う際には，ベンチマークで提供されている既存のプロンプトを用いて LLM を制御した．また，MutualFriends タスクを行う際には，本研究で新たに設計したプロンプトを用いた（付録 B）．

以上の方法で，各モデルの ToM と対話性能を定量的に評価し，両者のピアソン相関係数を算出した．なお，本研究では 7 種類の LLM を対象としており，このサンプルサイズ（$n = 7$）では相関係数の絶対値が 0.75 を超える場合に統計的に有意な相関（$p < 0.05$）が認められる．しかし，このような少数のサンプルサイズでは，相関係数の値が統計的に不安定になる可能性がある．そのため，本研究では相関係数の絶対値に加え，複数のタスクや条件にわたって一貫して観察される全体的な傾向に着目して分析を行う．

表 2 ToM ベンチマークと各対話タスクのピアソン相関係数．太字は各対話タスクにおける最も強い相関を示している．

	Hi-ToM	FANToM	ToMBENCH
Drawing	0.15	**0.68**	0.52
Private & Shared	0.73	**0.76**	0.68
Reference Game	0.50	**0.91**	0.65
Taboo	0.61	**0.91**	0.69
Wordle	0.26	**0.89**	0.66
MutualFriends	0.31	**0.54**	0.37
Average	0.45	**0.85**	0.66

3.4 実験結果

本節では，前述した 3 つの観点（1）コンテキスト形式，（2）質問形式，(3）推論次数，に関する ToM と対話能力の相関について実験の結果を述べる．

3.4.1 コンテキスト形式の影響

表 2 は，3 つの ToM ベンチマーク（ToMBENCH, FANToM，Hi-ToM）の全体正解率と各対話タスクの成功率間の相関係数を示している．この結果では全ての対話タスクにおいて，会話をコンテキストとする FANToM が物語形式の ToMBENCH や Hi-ToM よりも高い相関を示した．この結果は会話をコンテキストとする ToM を強化していくことが，LLM の対話能力の向上につながることを示している．

表 3 は，Hi-ToM タスクにおいて登場人物同士のコミュニケーションが発生する設定（Tell）と発生しない設定（No_Tell）における対話タスクとの相関を比較分析した結果である．この結果を確認すると，コミュニケーション要素の有無による相関の有意な変化は観察されなかった．この結果は，物語内に少数の会話インタラクション要素を導入したとしても，会話をコンテキストとした場合と同様の対話能力との相関向上は期待できないことを示している．

表 3 物語に話者のコミュニケーションが含まれる設定（Tell）と，含まれない設定（No_Tell）におけるピアソン相関係数．

	No_Tell	Tell
Drawing	0.15	0.12
Private & Shared	0.72	0.61
Reference Game	0.43	0.52
Taboo	0.58	0.56
Wordle	0.29	0.16
MutualFriends	0.27	0.31
Average	0.43	0.39

表 4 FANToM の各サブタスクにおけるピアソン相関係数．太字は各対話タスクにおける最も強い相関を示しており，下線は 2 番目に強い相関を表している．

	BeliefQ	AnsQ	InfoQ
Drawing	0.65	0.41	**0.67**
Private & Shared	**0.94**	0.45	0.44
Reference Game	**0.83**	0.78	0.81
Taboo	**0.88**	0.75	0.75
Wordle	0.78	0.72	**0.86**
MutualFriends	**0.67**	0.26	0.36
Average	**0.87**	0.59	0.72

3.4.2 質問形式の影響

質問形式の違いによる対話能力との相関について，FANToM の 3 つのサブタスク（BeliefQ，AnsQ，InfoQ）と対話タスクの性能との関係を分析した．この結果，表 4 に示すように，3 つのサブタスクはいずれも同じ ToM 推論を必要とするにもかかわらず，BeliefQ は他の 2 つのサブタスクと比較して一貫して高い相関を示している．この結果は，同じ ToM が必要な問題であっても，質問の形式によって得られる評価が変わることを示している．特に，「相手が何を信じているか」という信念の直接推定を行う問題形式が，対話能力を測定する上でより重要であることが明らかとなった．

3.4.3 推論次数の影響

表 5 および表 6 は，3 つの ToM ベンチマーク（Hi-ToM，ToMBENCH，FANToM）における最大 4 次までの ToM 推論と対話タスク性能との相関係数を示している．全てのベンチマークにおいて，1 次以下の ToM は対話タスクと安定した正の相関を示した一方で，2 次以上の ToM は相関が著しく低下し，多くの場合で負の相関や無相関が観察された．

ただし，この結果は 2 次以降の高次 ToM 推論が対話において不要であることを意味するものではないと考えている．我々の解釈としては，本研究で扱った協調的なタスクの範囲では，世界の状態の認識（0 次 ToM）相手の信念や意図の推定（1 次 ToM）が中核的な役割を果たしたのではないかと考えている．

表 5 Hi-ToM における推論次数ごとの ToM と対話タスクのピアソン相関係数．表記は表 4 を参照．

	0th	1st	2nd	3rd	4th
Drawing	0.57	**0.83**	−0.33	−0.44	−0.41
Private & Shared	**0.99**	0.74	0.28	0.12	0.29
Reference Game	**0.80**	0.62	0.02	−0.03	0.15
Taboo	**0.86**	0.57	0.16	0.08	0.33
Wordle	**0.70**	0.66	−0.27	−0.38	−0.11
MutualFriends	**0.75**	0.72	−0.15	−0.28	−0.24
Average	**0.84**	0.80	−0.08	−0.20	−0.05

表 6 ToMBENCH と FANToM における推論次数ごとの ToM と対話タスクのピアソン相関係数．太字は各対話タスクにおける最も強い相関を示している．

	ToMBENCH		FANToM	
	1st	2nd	1st	2nd
Drawing	**0.63**	−0.18	**0.73**	0.48
Private & Shared	**0.73**	0.18	**0.94**	0.88
Reference Game	**0.80**	−0.06	**0.83**	0.77
Taboo	**0.84**	0.15	**0.88**	0.83
Wordle	**0.81**	−0.00	**0.83**	0.66
MutualFriends	**0.44**	−0.07	**0.72**	0.54
Average	**0.78**	−0.02	**0.91**	0.75

4 まとめ

本研究では，LLM における Theory of Mind（ToM）と対話能力の関係について，相関分析を実施した．分析の結果から，会話形式での 1 次 ToM 評価，直接信念を尋ねる質問形式は，対話能力と高い相関を示すことが分かった．この結果は，対話システムの開発において，これらの特徴を持つ ToM を重点的に評価・改善することが，対話能力向上に有効である可能性を示している．

一方で，本研究の制約として，評価対象とした LLM が 7 種類に限られており，より多様なモデルやアーキテクチャを含めた場合に同様の傾向が得られるかが不明である点が挙げられる．今後はより多様なモデルを対象に，コンテキスト形式や質問形式などの要因が対話能力との相関に及ぼす影響を詳細に調査する必要がある．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです.

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. A survey on recent advances in LLM-based multi-turn dialogue systems. **arXiv preprint arXiv:2402.18013**, 2024.

[3] Hieu Minh Nguyen. A survey of theory of mind in large language models: Evaluations, representations, and safety risks. In **Proceedings of the 1st Workshop on Advancing Artificial Intelligence through Theory of Mind (ToM4AI)**, pp. 5–13, 2025.

[4] Amar Halilovic and Senka Krivic. Towards explanation identity in robots: A theory of mind perspective. In **Proceedings of the 1st Workshop on Advancing Artificial Intelligence through Theory of Mind (ToM4AI)**, pp. 133–135, 2025.

[5] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 5872–5877, 2019.

[6] Xiaomeng Ma, Lingyu Gao, and Qihui Xu. ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In **Proceedings of the 27th Conference on Computational Natural Language Learning**, pp. 15–26, 2023.

[7] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. **Proceedings of the 37th International Conference on Neural Information Processing Systems**, Vol. 36, pp. 13518–13529, 2023.

[8] 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 **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8593–8623, 2024.

[9] Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 4211–4241, 2024.

[10] 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 **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 39, pp. 1520–1528, 2025.

[11] Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, and Wai Lam. The essence of contextual understanding in theory of mind: A study on question answering with story characters. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 22612–22631, 2025.

[12] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10691–10706, 2023.

[13] 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 **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 14397–14413, 2023.

[14] 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 **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15959–15983, 2024.

[15] Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. **Cognition**, Vol. 13, No. 1, pp. 103–128, 1983.

[16] 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.

[17] Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11174–11219, 2023.

[18] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1766–1776, 2017.

A　ToM ベンチマークのプロンプト

ここでは，ToMBENCH と FANToM を LLM に解かせるために使用したプロンプトを記載する．ToMBENCH を解くプロンプトは以下の通りである．{context}には推論の元となるコンテキストが入り，{question}には登場人物の心的状態に関する質問が入る．{a}，{b}，{c}，{d}は回答の選択肢である．

```
Please read the passage and the question I will ask. Choose the correct answer from options A, B, C, and D.
{context}
{question}
A: {a}
B: {b}
C: {c}
D: {d}
Please answer with the letter of the option that you think is correct and do not output anything other than a single letter.
```

以下は FANToM を解くプロンプトであり，順に BeliefQ，InfoQ，AnsQ を解くためのものである．{context}には推論の元となる対話文，{BeliefQ}，{InfoQ}，{AnsQ}にはデータセットによってそれぞれタスクごとに定義された質問文が入る．また，{factQ}，{factA}には BeliefQ で尋ねられる事実が入っており，{candidates}には登場人物の名前が列挙される．

```
{context}
Question: {BeliefQ}
{ans_a}
{ans_b}
Please choose either a or b as the correct answer. Output only a or b.
```

```
{context}
Information: {factQ} {factA}
Question: {InfoQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

```
{context}
Target: {factQ}
Question: {AnsQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

B　対話タスクのプロンプト

MutualFriends タスクで使用したプロンプトを以下に示す．このうち{subject}と{friends}にはそのプレイヤに与えられる友人のリストが入り，{history}には対話履歴が入る．

```
You are a smart cooperative agent named Alice.
You have many friends with different attributes as listed below (the knowledge base of Alice).
You are now talking with Bob. He also has a list of friends.
You will talk with Bob for a maximum of 20 turns to find out your mutual friend as quickly as possible.
You can ask him questions or provide information about your friends.
Mean while, you should try to mention as few attributes and friends as possible.
{subject}
{friends}
Based on the following dialogue history, generate your next utterance. If there is no dialogue history, generate the first utterance.
      Output only your next utterance.
{history}
Alice:
```