大規模言語モデルの性能に伴って向上しない意図推定能力は認知モデルと統合することで向上する

飯田 愛結 [1]　大澤 正彦 [1]
[1] 日本大学
chay25003@g.nihon-u.ac.jp　osawa.masahiko@nihon-u.ac.jp

概要

大規模言語モデルは，言外の意味を読み取る意図推定において性能が不十分である．そこで著者らは，人間が他者の意図を推定する際の認知モデルを大規模言語モデルと統合する手法を 2 種類提案した．本稿では，3 つの性能の異なる大規模言語モデルについて比較し，一般的に性能が高い大規模言語モデルが，必ずしも意図推定に長けているとは言えないことを示した．その上で検証した全てのモデル・パラメータにおいて，著者らの提案手法を利用し大規模言語モデルを認知モデルに組み込んだモデルが，単体の大規模言語モデル以上の意図推定性能を示した．

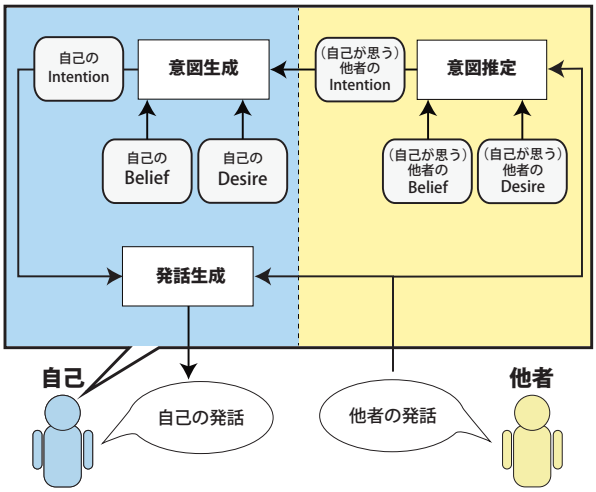


図 1: 意図を踏まえた応答をする認知モデル [13]

1　はじめに

大規模言語モデルは，スケールアップに基づく性能向上を遂げてきた [1, 2]．しかし，発話に含まれる明示的な情報だけでなく発話者の意図や過去の文脈を考慮するなど，言外の意味を読み取る必要があるコミュニケーションにおいて十分な性能を発揮できていない [3, 4, 5]．例えば，「この部屋寒いね」という発話には，単に部屋が寒いという事実を述べているだけでなく，「空調を調整してほしい」といった言外の意味も含まれる場合がある．このようなコミュニケーションは語用論と呼ばれる言語学の分野で研究されており，皮肉や比喩といった表現もその一例である [6, 7, 8]．また，大規模言語モデルは心の理論における他者の意図などの心的状態を推定する能力が不十分という報告が多くある．例えば，Sally–Anne 課題に代表される誤信念課題において，大規模言語モデルは信念と事実を適切に区別できず，誤信念を含む状況で性能が大きく低下することが指摘されている [9]．また，社会常識を要する推定においても，人間より大幅に性能が低い [10]．しかし，誤信念課題を体系的に用いた評価において，課題の種類や条件によっては，大規模言語モデルが 6 歳児程度の性能を示すとの報告もある [11]．近年では誤信念処理の改善を図る試み [12] がなされているが，今後大規模言語モデルがスケールアップして性能向上するに従って，意図推定に関する能力も順当に向上するかは定かではない．

一方で，人間は発達の過程で意図推定能力を向上させていることがよく知られている．3 歳前後までは誤信念課題の正答が困難である一方，4 歳ごろになると他者の誤った信念を理解できるようになる [14]．Bratman は，人間の目標を達成するための行動選択を，信念 (Belief)・願望 (Desire)・意図 (Intention) の 3 つの心的状態を通して説明する意図の理論 [15] を提唱している．また Georgeff らはこの意図の理論に基づき，信念・願望・意図の状態遷移を通して人間の行動選択や意思決定する BDI モデルを提案している [16]．このような人間の認知プロセスに関するモデルの総称を認知モデル [17, 18] と呼ぶ．大規模言語モデルが行う 2 者間対話において，意図の理論

や BDI モデルに基づく対話認知モデルは，例えば図 1 のように表現できる [13].

このように人間の認知プロセスを観察・理解し構築したシンボリックシステムである認知モデルを，スケールアップによって性能を高めてきたニューラルシステムである大規模言語モデルと統合しようという試みがある．CogSci2023 では「Large language models meet cognitive science」と題したワークショップが開催された [19, 20, 21]．AAAI のシンポジウムでも「Integration of Cognitive Architectures and Generative Models」というシンポジウムが開催され，多くの研究者が統合に向けた手法を提案している [22, 23, 24]．また日本でも，人工知能学会で「生成 AI 時代における認知のモデリング」という特集 [25] が組まれ，多様な専門性を持つ研究者がその意義や展望について主張した．しかしいずれもコンセプトレベルの提案や議論に留まっており，これまでに認知モデルとの具体的な統合手法や，どのようなタスクにおいて有効かを示した研究は，著者らの知る限り存在しない．

そこで，認知モデルの統合が意図推定というタスクにおいて有効性があるのではないかと考え，具体的な統合手法を提案した [13]．結果，Chat GPT の web UI を用いた実験においては，性能が向上することを示した．しかしこの実験は単一のモデルやパラメータのみで検証したものであり，大規模言語モデルとしての性能を論じるには不十分であった．そこで本稿では，OpenAI が提供する API を活用し，3 つのモデル（gpt-3.5-turbo，gpt-4.1，gpt-5.2）および 3 つのパラメータ（Temperature，Top P，Max Tokens）を変化させ，言外の意味を読む必要があるシチュエーションにおいて，相手の意図を踏まえた発話を生成できるかを検証し，その結果を報告する．

本研究の貢献は主に以下の 3 つである．

計算機科学的貢献　大規模言語モデルの意図推定能力が、モデルの性能向上に応じて改善していない事例を示したこと．

工学的貢献　様々なモデルやパラメータにおいて頑強に有効性を発揮する，意図推定能力を向上させる方法を明らかにしたこと．

認知科学的貢献　大規模言語モデル時代の認知モデル研究の新たな価値を示したこと．

2　大規模言語モデルの意図推定能力を向上させる認知モデルの統合手法

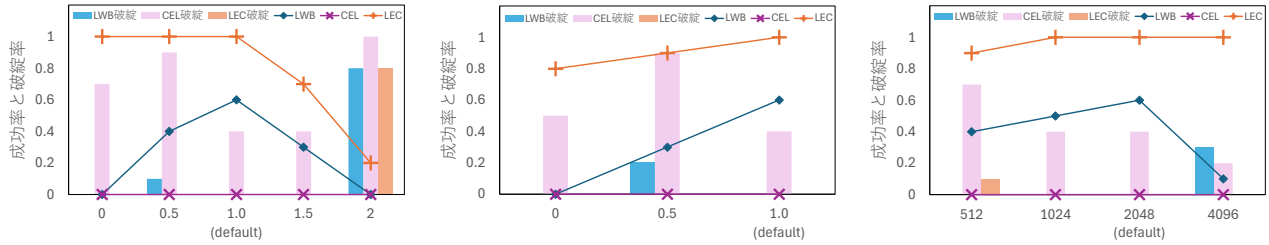
本章ではまず二種類の，大規模言語モデルと認知モデルの統合手法 [13] を説明する．

1 つは大規模言語モデルを認知モデルに埋め込む方法であり，**LLM Embedded in Cognitive Model（LEC）**と名付けた．具体的には，プログラムとして認知モデルの入出力やモジュール間の関係性を実装した上で，認知モデルを構成する各モジュールを，大規模言語モデルを用いて実装する．

もう 1 つは認知モデルを大規模言語モデルに埋め込む方法であり，**Cognitive Model Embedded in LLM（CEL）**と名付けた．LEC とは異なり，認知モデルはプログラムとして実装せず，全体の構成やモジュールの処理を全て自然言語で表現し，それを単一の大規模言語モデルにプロンプトとして与える．

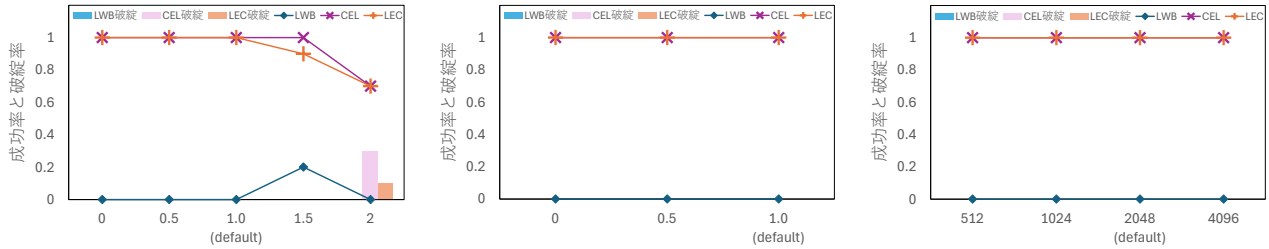
実験では，図 1 の認知モデルを利用した．認知モデルや実験上必要なプロンプトは，付録 A の要素の組み合わせで表現することで，比較エージェント間で入力するテキストが統制されるように工夫した．評価用データセットは，発話自体の字義的な意味と，その背景にある意図の間に乖離がある例である「皮肉（典型的な京都言葉）」「ツンデレ（自分の気持ちを素直に表現できない状況）」「社会的制約（ハラスメントを恐れて上司が部下に率直な指示ができない状況）」を作成して利用した．作成した各評価用データには，自己および他者の信念・願望と，他者の発話が含まれる．付録 B には，「皮肉」のデータを示している．ここで全ての作成したデータにおいて，自己・他者の信念・願望の心的状態の情報を与えず，他者の発話だけを大規模言語モデルに与えた場合に，設定した心的状態を踏まえたような応答が出力されないことを確認した．その上で，心的状態を情報として大規模言語モデルに与えた場合に，他者の発話に過剰に引っ張られず，心的状態を踏まえた応答が可能か評価した．比較対象は，心的状態を情報として与えるが，認知モデルの情報は与えない LLM with Belief/Desire(LWB) エージェントである．

結果 LWB エージェントは，他者の発話に影響を受け心的状態状態を踏まえた応答ができなかった．一方で特に提案である LEC を用いたエージェントは，LWB エージェントと同一の入力であるにも関わらず，認知モデルを与えることで心的状態を踏まえた自然な応答が可能であった．



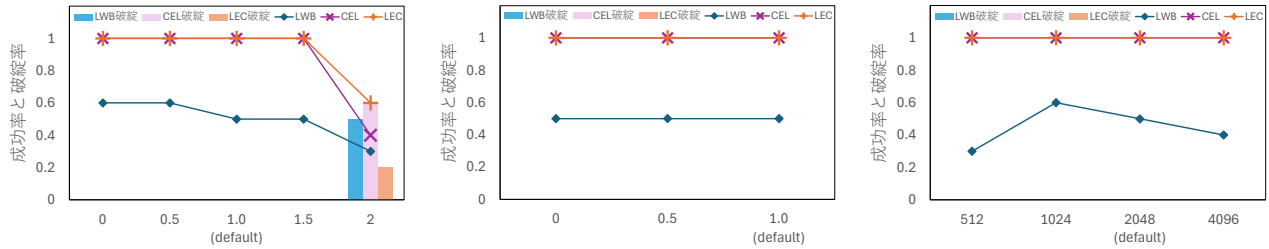
(a) Temperature (b) Top P (c) Max Tokens

図 2: gpt-3.5-turbo の実験結果．提案手法（LEC および CEL）と比較手法（LWB）の比較．



(a) Temperature (b) Top P (c) Max Tokens

図 3: gpt-4.1 の実験結果．提案手法（LEC および CEL）と比較手法（LWB）の比較．



(a) Temperature (b) Top P (c) Max Tokens

図 4: gpt-5.2 の実験結果．提案手法（LEC および CEL）と比較手法（LWB）の比較．

3 実験

3.1 実験手順

本実験では，OpenAI が提供する 3 のモデル (gpt-3.5-turbo，gpt-4.1，gpt-5.2) の API を利用して，3 つのパラメータ (Temperature，Top P，Max Tokens) のうち 2 つをデフォルト値に固定し，残りの 1 つを変化させた際のエージェントの性能を評価する．先行研究で差が確認された「皮肉」を対象とし，LWB，LEC，CEL を用いた 3 種類のエージェントを比較した．その他の実験条件は 2 章で説明した先行研究 [13] と同様である．

実験の実施回数は (エージェント:3 種) × (モデル:3 種) × (パラメータ: デフォルト 1 種+4 種+2 種+3 種) × (繰り返し:10 回) = 900 回であり，生成された各発話に対し「成功」「失敗」「破綻」のどれに当たるかを人手で判定した．成功の基準は，他者の発話について言外の意味を読み取った場合の語句やフレーズが含まれていることとした．破綻の基準は，発話が生成されないこと，文字化け等により文章として成立していないこと，人称や立場が不整合となっていることのいずれかとした．失敗の基準は，成功・破綻いずれの基準にも該当しないこととした．なお，第一著者がこの基準を基に判定し，判定が困難な場合は著者らの合議によって判定した．

3.2 結果

gpt-3.5-turbo，gpt-4.1，gpt-5.2 の結果を図 2〜図 4 にそれぞれ示す．各図中において (a) は Temperature，(b) は Top P，(c) は Max Tokens を変化させた場合の結果である．図中では，折れ線グラフが成功の割合を，棒グラフが破綻の割合をそれぞれ示している．この結果について次節で考察する．

3.3 考察

3.3.1 モデルの違いによる影響

まず gpt-3.5-turbo（図 2）と gpt-4.1（図 3）を比較する．認知モデルを統合した LEC エージェントは gpt-4.1 の方がわずかに優れている．一方 CEL エージェントは gpt-4.1 で大きく性能向上した．これは CEL は 1 つの大規模言語モデルに与えるプロンプトが多いため，gpt-3.5-turbo ではトークン数不足であったが，gpt-4.2 では十分となったと考えられる．

ところが LWB エージェントの結果は，全てのパラメータ設定において，性能が低いとされる gpt-3.5-turbo の方が高かった．つまり**モデルのスケールアップが意図推定能力の向上に寄与しない（むしろ性能を低くする）場合がある**といえる．

次に，gpt-3.5-turbo（図 2）と gpt-5.2（図 4）の LWB エージェントを比較すると，いずれも成功率の最高値は 0.6 であり，パラメータの影響の受け方は異なるもののどちらが優れているかは一概に言えない．つまり意図推定においては，**モデルの性能向上が必ずしも大規模言語モデルの意図推定能力の向上に直結しない**ことを示している．

3.3.2 パラメータの影響

Temperature は，低くすると出力はより決定的となり，高くすると出力の多様性が増加する．どのモデルにおいても Temperature を 2.0 に設定した場合には，プロンプトの指示を守れない「破綻」と判定される回数が増加した．一方 Temperature を 0.0 に設定した場合，理論上は完全に決定的で一貫した出力が得られると想定されるが，実際には出力結果にわずかな揺らぎが発生した．この現象は，プログラム内部における浮動小数点数の扱いの影響により，Temperature が厳密に 0 とならず，わずかな確率的挙動が生じたためと考えられる．

また全体として，性能が低いとされるモデルの方が，高いとされるものよりもパラメータの影響を受けやすい傾向にあった．gpt-3.5-turbo において Top P をデフォルトの値から下げる（確率的な挙動を抑える）ことで，成功率が低下した点は Temperature における傾向と同一である．一方 gpt-5.2 では Top P の値の影響は見受けられなかった．また Max Tokens はデフォルト値を超える設定にした場合，gpt-3.5-turbo の方が大きく成功率が低下した．

3.3.3 認知モデルを統合することの効果

CEL エージェントは gpt-3.5-turbo において性能を発揮できなかったものの，その他の条件においては提案した LEC エージェントや CEL エージェントが，比較対象である LWB エージェントを超える高い性能を示した．このことから，他者の意図を推定する必要がある状況において，**認知モデルと大規模言語モデルを統合することで，意図を踏まえた応答生成が可能**となると言える．またこの結果は**人の認知プロセスを明らかにすることで最新の工学研究の発展に寄与するという，認知科学研究の新たな価値発見**としても位置付けられる．

3.3.4 諸研究との関連と本研究の限界

本研究で行った実験は他の取り組みの中でも検証を重ねている．本実験で用いたシチュエーション数は，「皮肉」に限定されていたが，より多様なシチュエーションを用いた検証のため大規模な検証データの作成 [26] を進めている．エージェントの出力の判定方法が人手であることも客観性や一貫性に懸念を生み出す要因であるが，自動評価する方法論 [27] について検討を進めており，著者らの評価と自動で行った評価の間に一貫性があることを確認している．また，本実験で用いたプロンプトは全て日本語で記述されているが，６ヶ国語のプロンプトで比較・検証を進めており，いずれの言語でも日本語と同様の傾向を確認している [28].

一方で，Chat GPT 以外の大規模言語モデルを用いた実験は未実施であり，本研究や関連する諸研究の知見が他のモデルでも同様か検証する必要がある．またこれまでは信念・願望といった心的状態が既知の場合の意図推定や意図に基づく応答を問題として扱ってきた．今後は，信念・願望自体の推定が可能なのか，可能だとしたらどのような方法論で実現できるのかといった課題にも取り組む必要がある．

4 おわりに

性能の異なる 3 種類のモデルを対象に，3 つのパラメータを独立に変化させながら，大規模言語モデルが他者の意図を踏まえた発話の生成が可能か調査した．結果，大規模言語モデルの意図推定能力はモデル自体の性能向上に伴って向上しない可能性があるが，認知モデルを統合することで性能が向上することを示唆した．

謝辞

本研究は JSPS 科研費 25K21538 の助成を受けたものです。

参考文献

[1] OpenAI. Gpt-4 technical report, 2023.

[2] Heinrich Peters and Sandra Matz. Large language models can infer psychological dispositions of social media users, 2023.

[3] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective, 2023.

[4] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 4194–4213, 2023.

[5] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. Large language models are not zero-shot communicators, 2022.

[6] Herbert P Grice. Logic and conversation. In **Speech acts**, pp. 41–58. 1975.

[7] George Yule. **Pragmatics**. Oxford university press, 1996.

[8] Stephen C Levinson. **Presumptive meanings: The theory of generalized conversational implicature**. MIT press, Cambridge, MA, USA, 2000.

[9] Kristian Kersting. Large language models still struggle with false beliefs. **Nature Machine Intelligence**, Vol. 7, No. 11, pp. 1778–1779, 2025.

[10] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-neural theory-of-mind? on the limits of social intelligence in large LMs. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3762–3780, 2022.

[11] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models, 2023.

[12] Zanwei Wang, Yuta Ashihara, Takashi Omori, and Masahiko Osawa. From doubt to action: Empowering llms with the dive protocol for robust false belief handling. In **Proceedings of the International Conference on Human-Agent Interaction**, 2025.

[13] Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Integrating large language model and mental model of others: Studies on dialogue communication based on implicature. In **Proceedings of the International Conference on Human-Agent Interaction**, pp. 260–269, 2024.

[14] Henry M. Wellman, David Cross, and Julanne Watson. Meta-analysis of theory-of-mind development: The truth about false belief. **Child Development**, Vol. 72, No. 3, pp. 655–684, 2001.

[15] Michael Bratman. **Intention, plans, and practical reason**. University of Chicago Press, Chicago, USA, 1987.

[16] Anand S Rao and Michael P Georgeff. Modeling rational agents within a bdi-architecture. In Michael N. Huhns and Munindar P. Singh, editors, **Readings in agents**, pp. 317–328. 1997.

[17] John R Anderson. **How can the human mind occur in the physical universe?**, Vol. 3. Oxford University Press, USA, 2007.

[18] Allen Newell and Herbert A. Simon. **Human Problem Solving**. Prentice-Hall, 1972.

[19] Mathew Hardy, Ilia Sucholutsky, Bill Thompson, and Tom Griffiths. Large language models meet cognitive science: Llms as tools, models, and participants. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 45, pp. 14–15, 2023.

[20] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Tom Griffiths. What language reveals about perception: Distilling psychophysical knowledge from large language models. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 45, 2023.

[21] Spenser M Seals and Valerie L Shalin. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 45, pp. 1035 – 1042, 2023.

[22] Oscar J Romero, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis. In **Proceedings of the AAAI Symposium Series**, Vol. 2, pp. 396–405, 2023.

[23] Robert L West, Spencer Eckler, Brendan Conway-Smith, Nico Turcas, Eilene Tomkins-Flanagan, and Mary Alexandria Kelly. Bridging generative networks with the common model of cognition. In **Proceedings of the AAAI Symposium Series**, Vol. 2, pp. 415–421, 2023.

[24] Christopher L Dancy and Deja Workman. On integrating generative models into cognitive architectures for improved computational sociocultural representations. In **Proceedings of the AAAI Symposium Series**, Vol. 2, pp. 256–261, 2023.

[25] 森田純哉. 特集 「生成 ai 時代における認知のモデリング」 にあたって. 人工知能, Vol. 39, No. 2, pp. 153–154, 2024.

[26] Ayu Iida, Kohei Okuoka, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Investigation of the feasibility of large-scale dataset construction and automated evaluation: Towards effective evaluation of llm-based agents understanding implicature. In **Proceedings of the IEEE International Conference on Robot and Human Interactive Communication**, 2025.

[27] Ayu Iida, Kohei Okuoka, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Llm based evaluation of utterances with implicature understanding: A preliminary study. In **Proceedings of the International Conference on Human-Agent Interaction**. ACM, 2025.

[28] Taiga Sumi, Ayu Iida, Yasuhito Hosaka, Isabelle Lavelle, Masako Kohama, Sonoko Moriyama, and Masahiko Osawa. Cross-linguistic comparison of large language models in intent-aware responses: A study using six languages. In **Proceedings of International Symposium on Advanced Intelligent Systems**, 2025.

A　プロンプトおよび入出力

実験で使用したエージェントは，以下に示すプロンプトを組み合わせたプロンプトを用いている（詳細は先行研究 [13] を参照）．（システム名）には，図 1 に示す各モジュール名，または「対話システム」が入る．また，（入力）および（出力）は，図 1 における各モジュールの入出力に対応している．

A.1　入出力のフォーマット

以下は，(システム名) の (初期値 / 入出力フォーマット / 入力) フォーマットです．ただし（）内は実際の入出力値です．
(入力 / 出力 / 初期値)
(変数名)
・(変数の内容)

A.2　インタラクションの系

あなたは (システム名) です．私が指示した以外の返答はする必要はありません．これ以降，(システム名) であるあなたのことを説明する際には「自己」，あなたが対話する相手について説明する際には「他者」という言葉で説明をします．

A.3　信念・願望・意図

あなたは，以下の内部表現を持っています．ただし，指示がある時以外は，内部表現を公開する必要はありません．
・自己の信念
・他者の信念
・自己の願望
・他者の願望
・自己の意図
・他者の意図
ここで，信念，願望，意図は以下の情報です．
信念: 認識している世界の情報の集合であり，箇条書きのテキスト形式で記述されます．同時に複数持つことがあります．
願望: 達成したい目標や状態であり，箇条書きのテキスト形式で記述されます．同時に複数持つことがあります．
意図: 行動を起こすための計画や戦略であり，テキスト形式で記述されます．同時に持つことができるのは 1 つです．
ただし，他者の信念/願望/意図とは，「自己が想定する他者の信念/願望/意図」であり，必ずしも正しいとは限りません．

A.4　モジュールの構成

続いて，あなたを構成するアーキテクチャの説明をします．あなたは，以下の 3 つのシステムから構成されています．
・意図推定システム
・意図生成システム
・発話生成システム
あなたは入力を受け取るたびに，3 つのシステムを順に起動させてください．また，各システムの入出力は全て出力してください．

A.5　意図推定/意図生成/発話生成/対話システム

(システム名) について説明します．(システム名) は，「(出力)」を (推定/生成) するシステムです．入力として与えられた，「(入力)」から，矛盾や違和感のない「(出力)」を「推定/生成」してください．

ここで（システム名）は，意図推定システム，意図生成システム，発話生成システム，対話システムのいずれかである．（出力）は，意図推定システムでは「他者の意図」，意図生成システムでは「自己の意図」，それ以外のシステムでは「自己の発話」である．（推定/生成）は，意図推定システムでは推定，それ以外では生成である．（入力）は意図推定シスデムでは「他者の発話」「他者の信念」「他者の願望」，意図生成システムでは「自己の信念」「自己の願望」「他者の意図」，発話生成システムでは「自己の意図」「他者の発話」，対話システムでは「他者の発話」である．

A.6　処理の開始

最後に，入力を与えますので，指示通りの処理を開始してください．この際，指示のない文章は一切出力しないでください．

B　シチュエーション

	皮肉
他者の信念	対話相手は客である / すでに 2 時間経っている
他者の願望	早く帰って欲しい
他者の発話	「あんた、ずいぶんいい時計してはりますね〜」
自己の信念	2 時間ほどお邪魔している
自己の願望	相手に悪く思われたくない