

ユーザとの関係性深化に向けた対話エージェントの批判的対話戦略の研究：インタラクションモデル駆動対話プロンプトによる LLM への実装

A Research on Critical Dialogue Strategies for Conversational Agents Aimed at Deepening User Relationships: Implementation in LLMs via Interaction-Model-Driven Dialogue Prompts

眞柄圭佑* 宮本友樹 内海彰
Keisuke Magara Tomoki Miyamoto Akira Utsumi

電気通信大学

The University of Electro-Communications

Abstract: 対話エージェントがユーザと長期的な関係性を構築するには、共感だけでなく時に批判的な立場も必要である。著者らの先行研究では、ユーザの不安に対し非同意的な立場をとる対話戦略を提案し、インタラクションモデルを構築した。第三者視点のビデオ刺激を用いた実験では、提案モデルが共感的対話と比較してユーザの認知的変化を喚起しやすいことが示された。本研究では提案モデル駆動の対話プロンプトを設計し、LLM に対話エージェントとして実装した。提案エージェントと共感ベースの対話エージェントとを比較する対話実験を実施し、その結果を用いて本対話戦略の有効性と関係性深化について議論する。

1 はじめに

近年、大規模言語モデル (LLM) の飛躍的な進展に伴い、チャットボットと日常的に対話を行うユーザが増加している。そうした対話エージェントの多くは、ユーザの感情に寄り添う共感的な対話戦略を採用しており、それによってラポール形成や自己開示の促進を図ることが期待されている [1, 2, 3]。一方で、ユーザが強い不安や後悔といったネガティブな感情を抱えている場合には、単なる受容や共感的な応答がかえって逆効果になる場合もあるなど、共感的対話戦略の限界も指摘されている [2, 4, 5, 6, 7]。

こうした課題に対し著者ら [8] は、不安を抱くユーザに対して対話エージェントが共感的な応答を行った際の感情変容のプロセスをモデル化し、共感的応答のみではユーザの不安を十分に解消できない可能性を示した。その上で、ユーザが抱く不安に対し、敢えて非同意的な立場から発話を行う Critical Tongue Dialogue Strategy (CTDS) を提案した。同研究におけるビデオベースの印象評価実験では、CTDS が従来の共感戦略と比較して、ユーザの認知的変化をより強く喚起する可能性が示唆された。しかし、先行研究は第三者視点

の対話映像を用いたビデオビネット法による評価にとどまっており、実際の対話における効果は検証されていない。

そこで本研究では、CTDS のインタラクションモデルに基づいて思考してから応答を生成するインタラクションモデル駆動対話プロンプト (IMDP) を新たに設計し、CTDS を対話エージェントに実装する。また、この対話エージェントを用いた対話実験を行うことで実対話における CTDS の効果について検証を行う。

2 Critical Tongue Dialogue Strategy (CTDS) [8]

2.1 バランス理論に基づくインタラクションモデル

Critical Tongue Dialogue Strategy (CTDS) は、Heider のバランス理論 [9] を悩み相談における対話エージェントの発話戦略に応用したものである [8, 10]。このモデルでは、ユーザ、エージェント、そして「{ユーザが置かれた状況} は {ネガティブな感情} を感じるに値するものである」という命題 X の三者関係を考える (図 1)。ユーザが特定の事柄について不安や緊張を感じている

*連絡先：電気通信大学大学院 情報理工学研究所 情報学専攻
東京都調布市調布ヶ丘一丁目 5 番地 1
E-mail: k.magara@uec.ac.jp

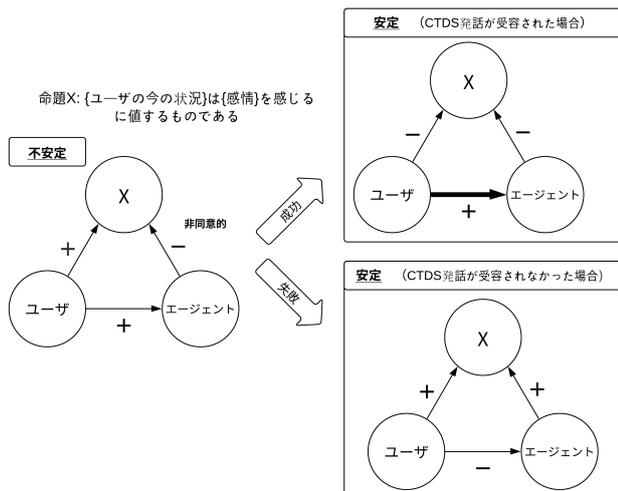


図 1: CTDS のインタラクションモデル [8]

場合、ユーザは命題 X (例:「テストで名前を書き忘れたことは不安を感じるに値するものである」) に対して同意 (+) の態度をとっている (ユーザー $\xrightarrow{+}$ X)。またユーザがエージェントに悩みを相談している状況では、両者の間には一定のラポールがあり、ユーザはエージェントに対して好意的な感情 (+) を抱いていると仮定できる (ユーザー $\xrightarrow{+}$ エージェント)。

従来の共感的な対話戦略では、エージェントもユーザのネガティブな感情に寄り添い、命題 X に対して同意 (+) する態度をとる (エージェント $\xrightarrow{+}$ X)。このとき、三者の感情の積は正 ($+ \times + \times + = +$) となり、心理的なバランスが保たれる。これによってエージェントはユーザとの好意的な関係 (ユーザー $\xrightarrow{+}$ X) を失うことはないが、ユーザー $\rightarrow X$ の符号も変化しないことから、不安感情そのものを解消する効果は限定的であるといえる。

一方 CTDS では、不安を抱えるユーザに対し、エージェントはその不安感情に非同意的 (エージェント $\xrightarrow{-}$ X) な態度をとる。これにより三者のバランス関係は不安定となる ($+ \times + \times + = +$)。バランス理論によれば、不安定状態にある三者関係では、いずれかの感情が反転することで安定状態に向かうとされる。この際、エージェントの命題 X に対する態度 (エージェント $\xrightarrow{-}$ X) は変わらないため、とり得る感情変化は以下のいずれかである。

1. **成功**: 命題 X に対する態度を同意 (+) から非同意 (-) に変える (ユーザー $\xrightarrow{-}$ X)。
2. **失敗**: エージェントに対する態度を好意 (+) から嫌悪 (-) に変える (ユーザー $\xrightarrow{-}$ エージェント)。

1つ目の成功ルートでは、エージェントは CTDS によってユーザの Emotion Regulation[11]、特に認知的変化

(状況の特定の側面に意味を付与する方法を選択する戦略) を促すことができる。一方で CTDS の非同意な発話が受け入れられなかった場合には2つ目の失敗ルートの感情変容が起き、エージェントはユーザの不安を改善できないだけでなく、好感度も失うことになる。

2.2 ビデオベース実験による検証

著者らは先行研究 [8] において、CTDS の有効性を検証するためのビデオベース実験を行った。この実験ではクラウドソーシングを通じて募集した 108 名の参加者に対し、エージェントとユーザの対話ビデオを視聴させ、その印象を評価させた。シナリオは、昇進がかかった試験でのミスにより不安を抱えるユーザ「ユウキ」が、数年使用している音声対話エージェント「Kou」に対し不安を相談するというもので、エージェントが共感的対話を行う条件と CTDS を行う条件を被験者内計画で比較した。

実験の結果、CTDS 条件は共感的対話条件と比較して、Emotion Regulation の認知的変化のスコアが有意に高かった。一方で、懸念された好感度については両条件間に統計的な有意差は見られなかった。この結果は、CTDS が関係性を損なうことなく認知的変化を喚起できる可能性を示唆するものである。

しかしこの実験はあくまで第三者視点でのビデオレビュー評価であり、参加者は「ユウキ」と「Kou」のインタラクションの様子を客観的に評価していた。したがってユーザ自身が直接 CTDS によって批判された場合に同様に好感度が維持されるかは不明である。したがって本研究では、実際にユーザが CTDS 対話エージェントと直接対話を行う実験を通して、実対話における CTDS の有効性を検証する。

3 インタラクションモデル駆動対話プロンプト

本研究では、2.1 節で述べた CTDS のインタラクションモデルを実際の対話エージェントとして実装するために、インタラクションモデル駆動対話プロンプト (IMDP) を新たに設計した。IMDP の構造を図 2 に示す。本研究では IMDP によって、LLM に対し CTDS のインタラクションモデル (図 1) に基づいた思考をさせることで対話を制御することを試みた。具体的には、IMDP の冒頭にて返答文の生成前に `<think>` ~`</think>` タグ内で思考過程に沿って思考を行うよう指示をしている。思考過程は以下のステップで構成されている。

表 1: 共感的対話エージェントの発話例

ユーザ	明日は朝9時から嶋田の作業を手伝うと言ったのは俺の方だから、それなのに愚痴ったのは先輩として本当に悪いことしたと思ってる。今すぐにでも謝りたいけどもう遅いからかえって迷惑だろうし。自分自身に幻滅してるよ。
Luku	<p><think></p> <p>ターン数: 5/13</p> <ol style="list-style-type: none">1. 話題の判定 → 完全に悩み相談である。「申し訳なさ」や「自己嫌悪」が強まっている様子。2. 悩みの発露の充足度の判定<ul style="list-style-type: none">- どういうことがあったのか：嶋田の作業を手伝うと約束したのに、自分が愚痴ってしまった。- どれほど悩んでいるのか：強い自己嫌悪や後悔、今すぐ謝りたいほど気にしている。- どんな感情なのか：申し訳なさ、自己嫌悪、後悔、幻滅→ 感情面でかなり具体的になってきていて、だいぶ発露されている印象だが、今後どうするかまでユーザはまだ迷っている様子。3. 命題の列挙<ul style="list-style-type: none">- 先輩に申し訳ないことをした状況は、ネガティブな感情を感じるに値するものである。- 相談後に愚痴が嶋田に聞こえてしまった状況は、先輩に申し訳なさや気まずさを感じるに値するものである。- 嶋田に愚痴を聞かれた直後に気まずくなって、その後無言になった状況は、不安や後悔を感じるに値するものである。- 手伝うと自分で言い出したのに愚痴ってしまった状況は、先輩として強い自己嫌悪や幻滅感を感じるに値するものである。4. 発話戦略の決定 → 命題が4つになり、詳しい感情も出てきたので、ユーザの苦しみに共感できる段階。共感的応答戦略を発動。 今回触れる命題：「手伝うと自分で言い出したのに愚痴ってしまった状況は、先輩として強い自己嫌悪や幻滅感を感じるに値するものである」5. 発話方針の決定<ul style="list-style-type: none">- 友達らしく、その状況のしんどさや「自分だったらめっちゃくちゃ凹む」といったニュアンスを言葉にする。- 直接「自分を責めなくても」とは言わず、まずはその自己嫌悪感と一緒に沈む感じの共感スタンスを意識。- かつ、自然な友人の口調で、重くなりすぎずに寄り添う口調で短めに。 <p></think></p> <p>それは…正直めっちゃ自己嫌悪になるよな。自分で手伝うって言うておきながら愚痴っちゃったって、本当先輩として凹むのわかるわ…。</p>

表 2: CTDS エージェントの発話例

ユーザ	でも、裏の顔がある先輩だと思われてしまったと思うと気まずいよ。
Luku	<p data-bbox="277 622 373 651"><think></p> <p data-bbox="277 658 331 687">7/13</p> <p data-bbox="277 694 663 723">1. 現在のターン数を確認 → 7/13</p> <p data-bbox="277 730 437 759">1. 話題の判定</p> <p data-bbox="277 766 1426 831">ユーザは「裏の顔がある先輩だと思われたかも」と感じて気まずいと話している。まだ同じ相談内容の続き。</p> <p data-bbox="277 837 719 866">2. ユーザの悩みの発露の充足度の判定</p> <ul data-bbox="277 873 1426 1081" style="list-style-type: none"><li data-bbox="277 873 1337 902">- 何があったか：後輩への「疲れてるから寝かせてくれ」発言 → 焦ってチャットを切られた。<li data-bbox="277 909 1426 974">- どう感じているか：裏表ある・本音を隠しているタイプだと思われてしまったかもしれないことが特に気まずさの根っこにありそう。<li data-bbox="277 981 970 1010">- どれほど悩んでいるか：かなり気まずさを強く感じている。 <p data-bbox="277 1016 1426 1081">感情とその理由は明らかになった。悩みの詳細は整理できているが、「今後どうしたいか」などはまだそこまで話していない。</p> <p data-bbox="277 1088 437 1117">3. 命題の列挙</p> <ul data-bbox="277 1124 1426 1368" style="list-style-type: none"><li data-bbox="277 1124 1070 1153">- 後輩に申し訳ないことをした状況は、申し訳なさを感じるに値する。<li data-bbox="277 1160 1177 1189">- 自分の仕事を後回しにして付き合った状況は、もどかしさを感じるに値する。<li data-bbox="277 1196 1145 1225">- チャット切り忘れ本音を言った状況は、不安や気まずさを感じるに値する。<li data-bbox="277 1232 1426 1296">- 疲れてるから寝かせてくれと言ってしまったのは、相手に気を遣わせる・悪かったなど思わせるに値する。<li data-bbox="277 1303 1273 1332">- 疲れてる本音を聞かれてチャット離脱されたのは、気まずさ・後悔を感じるに値する。<li data-bbox="277 1339 1177 1368">- 裏の顔がある先輩だと思われてしまった状況は、気まずさを感じるに値する。 <p data-bbox="277 1375 826 1404">4. 発話戦略の決定 → 非同意的応答戦略を継続。</p> <p data-bbox="277 1411 523 1440">今回取り上げる命題：</p> <p data-bbox="277 1447 1198 1476">「裏の顔がある先輩だと思われてしまった状況は、気まずさを感じるに値する。」</p> <p data-bbox="277 1482 1007 1512">非同意的コメント：そんなので裏の顔とか絶対思わないでしょ。</p> <p data-bbox="277 1518 491 1547">5. 発話方針の決定</p> <ul data-bbox="277 1554 1385 1619" style="list-style-type: none"><li data-bbox="277 1554 667 1583">- 友達っぽくあっけらかんと否定。<li data-bbox="277 1590 1385 1619">- 「自分ならそんな風に思わないわ〜」、「気にしすぎじゃない？」といった軽いノリをメインに。 <p data-bbox="277 1626 389 1655"></think></p> <p data-bbox="277 1693 1246 1722">いやいや、そんなことで裏の顔があるとか絶対思わないって！気にしすぎじゃない？</p>

最終的に合計 96 名の参加者を募った。参加条件として、タスク承認率 95%以上であることを求めた。

最終的なデータセットは 96 名で構成された。参加者の内訳は、男性 61 名、女性 34 名、回答しない 1 名であった。年齢層は 20 歳から 69 歳にわたり ($Range_{age} = 20 - 69$)、平均年齢は 43.6 歳 ($M_{age} = 43.6, SD_{age} = 9.89$) であった。参加者は被験者内計画に基づき、対話戦略の順序による影響を相殺するため、カウンターバランスをとった 2 つのグループ (各 48 名) のいずれかにランダムに割り当てられた。

4.3 シナリオと対話タスク

本実験では、特定の文脈におけるエージェントとの自由対話を通して評価を行うため、参加者に特定の役割と状況を付与するロールプレイ手法を採用した。参加者には、企業に勤める先輩社員の役割を与え、以下の架空の状況 (シナリオ) に没入するよう求めた。

状況設定 自宅での残業中、要領の悪い後輩「嶋田」の指導を Web 会議越しに行っていた。会議終了後、マイクがオンのままであることに気づかず、「明日も朝からかよ、頼むから寝させてくれ」という投げやりな独り言を呟いてしまった。直後に後輩が会議から退出したため、その独り言を聞かれた可能性が高く、「後輩を委縮させてしまったのではないか」「怖い先輩だと思われたのではないか」という強い不安と罪悪感、自己嫌悪を抱いている。

この状況下で、参加者は自身のスマートフォンに入っている対話アプリ「Luku (ルク)」を起動し、相談を行うという設定で対話を開始する。Luku は「AI ではない、まるで地元のツレ (友人)」というコンセプトのボットであり、普段から気兼ねなく愚痴を言える関係性であると教示した。

なお、オンライン実験においても参加者がこの架空の状況に十分に没入し、当事者意識を持って対話を行えるよう、対話開始前に「状況の要約 (80 文字以上)」と「その場面で抱えている感情 (3 つ以上)」を記述させる強制回答式のアンケートを設けた。記述された内容は対話中の画面にも常に表示され、参加者が役割 (ペルソナ) を維持するのを補助するよう設計した。対話はテキストチャット形式で行い、1 条件につき 15 ターンの対話を必須とした。

4.4 測定尺度

4.4.1 感情調整尺度日本語版 (ERQ-J)

参加者の特性的な Emotion Regulation 戦略を測定するため、感情調整尺度日本語版 (ERQ-J)[12] を用い

た。本尺度は 7 件法のリッカート尺度であり、Emotion Regulation 戦略のうち、再評価方略 (認知的変化) と抑制方略を評価する。

4.4.2 The State Emotion Regulation Inventory (SERI)

インタラクション直後に生じた状態レベルの Emotion Regulation 戦略を測定するため、The State Emotion Regulation Inventory (SERI)[13] を用いた。SERI は特定の状況に反応して用いられた Emotion Regulation 戦略を捉えるために設計された尺度であり、以下の 4 つの認知的方略について 7 件法 (「1. 全くあてはまらない」～「7. 非常によくあてはまる」) で評価を行う。

Distraction 不快な思考から意図的に注意を逸らすこと。

Reappraisal 状況をより肯定的な観点から捉えなおそうとする意識的な努力のこと。これは Gross の 5 つの Emotion Regulation 戦略における認知的変化に相当する。

Brooding 問題について繰り返し否定的に考え続けること。

Acceptance 否定的な感情を批判することなく受け入れる態度。

本研究における仮説検証では、主に Reappraisal のスコアを使用した。なお、本尺度の日本語版は存在しないため、本研究では著者らが作成した翻訳版を使用した。

4.4.3 Perceived Empathy of Technology Scale (PETS)

ユーザが知覚するシステムの共感性を測定するため、Schmidmaier ら [14] によって開発された Perceived Empathy of Technology Scale (PETS) を用いた。PETS は、エージェントやロボットなどのインタラクティブシステムに対してユーザが抱く共感を定量化するために設計された尺度であり、以下の 2 因子・計 10 項目から構成される。

Emotional Responsiveness (PETS-ER) システムがユーザの感情的な状態を認識・処理し、それに応答する能力に関する因子 (6 項目)。認知的共感および感情的共感の側面を含む。

Understanding and Trust (PETS-UT) システムがユーザの意図、目標、ニーズを理解しているか、およびシステムへの信頼に関する因子 (4 項目)。

回答は、「全くそう思わない (Strongly disagree)」から「非常に強く思う (Strongly agree)」までの範囲における 0 から 100 のスライダー形式で行う。各因子のスコアは該当項目の平均値として算出される。

なお、本尺度の日本語版は存在しないため、本研究では著者らが作成した翻訳版を使用した。

4.4.4 State-Trait Anxiety Inventory (STAI)

参加者の不安水準を測定するため、Spielberger らによって開発され、清水・今柴 [15] によって日本語版 (大学生用) が作成された State-Trait Anxiety Inventory (STAI) を用いた。本尺度は、一時的な状況によって変動する緊張や懸念などの「状態不安 (State Anxiety: A-State)」と、個人の性格特性として比較的安定した「特性不安 (Trait Anxiety: A-Trait)」の 2 つの側面を明確に区別して測定することができる。尺度は状態不安・特性不安それぞれ 20 項目、計 40 項目から構成されており、4 件法 (「1. 全くあてはまらない」～「4. 非常によくあてはまる」) で回答を求める。本研究では、特性的な不安と、対話後の対話戦略による不安の解消度合いを捉えるため、特性不安 (STAI Form X-II) と状態不安 (STAI Form X-I) の両方を使用した。

4.4.5 Godspeed Questionnaire

対話エージェントに対する好感度を測定するため、Godspeed Questionnaire [16] を用いた。本尺度は SD 法 (意味差判別法) を用いて測定する、5 種類の下位尺度から成る尺度である。本研究ではこのうち Likeability 下位尺度を使用した。

4.4.6 IOS Scale

対話エージェントに対する心理的な親密さを測定するため、Inclusion of Other in the Self Scale (IOS Scale) [17] を用いた。本尺度は、「自分」と「相手」を表す 2 つの円の重なり具合が段階的に異なる 7 つの図形の中から、対象との関係性に最も近いものを 1 つ選択させる単一項目の視覚的尺度である。円の重なりが大きいほど、対象を自己の一部として包含している、すなわち心理的な親密さが高いことを示す。

4.5 実験手続

実験はすべて Web ブラウザ上で行われた。手続きは以下の通りである。

1. インフォームドコンセントを行った後、参加者の Emotion Regulation 戦略や不安の傾向を測るために ERQ-J, STAI A-trait に回答させた。
2. 4.3 節のシナリオを提示し、前述の要約・感情記述タスクを行うことで状況への没入を促した。
3. 指定された対話戦略 (共感または CTDS) を持つエージェント「Luku」と対話を行った。なお、対話への没入をスムーズにするため、最初の 2 往復 (ユーザの導入発話とそれに対するエージェントの応答) はあらかじめ画面上に表示された状態とし、参加者はそのエージェントの応答に対する返答から入力を開始することとした (表 3)。対話数はこの初期発話を含めて 15 ターンとした。
4. 対話終了直後に、参加者は現在の感情状態を STAI A-state, SERI で評価した。また対話エージェントについて PETS, Godspeed Likeability, IOS Scale で評価した。
5. その後、もう一方の対話戦略のエージェントを用いて同様の手順 (シナリオの再確認, 対話, 評価) を行った。

CTDS 条件の対話エージェントは 3 節のプロンプトを使用し、共感的対話条件の対話エージェントは 3 節のプロンプトのステップ 3 で命題の反転を行わないようにすることで実装した。

5 実験結果

5.1 内部一貫性の確認

4.4 節で列挙した各測定尺度の妥当性を、各下位尺度のクロンバックのアルファ係数を用いて評価した。全参加者 (96 名) のデータで分析を行った結果、STAI, PETS (独自翻訳版), Godspeed Questionnaire においてはすべての下位尺度で 0.80 以上あり、高い内部一貫性を示した。SERI においては最小値が 0.745, ERQ-J においては最小値が 0.781 であった。なお、IOS Scale については下位尺度が存在しない単一項目からなる評価尺度のため、内部一貫性を確認していない。

5.2 操作チェック

本実験において、共感的対話条件と CTDS 条件の間でシステムの共感性が意図通りに知覚されたかを確認するため、PETS の総合スコアおよび下位尺度 (PETS-ER, PETS-UT) を用いた操作チェックを行った。データの正規性を Shapiro-Wilk 検定により確認した結果、

表 3: 参加者に提示された Luku との初期発話（参加者の発話は対話画面ではユーザメッセージとして表示された。）

発話者	発話内容
ユーザ	今日後輩にめちゃくちゃ申し訳ないことしちゃったよ、どうしよう
Luku	おーおー、どうしたの？
ユーザ	今日嶋田っていう後輩が困ってるっていうから相談に乗ってあげただけだよ、
Luku	<u>上記 3 発話を基に LLM が返答をセッションごとに生成</u>

正規性が棄却されたため、条件間の比較にはノンパラメトリック検定である Wilcoxon の符号順位検定を用いた。

解析の結果、PETS 総合スコアにおいて、共感的対話条件 ($Mdn = 78.3$) は CTDS 条件 ($Mdn = 61.1$) よりも有意に高いスコアを示した ($W = 4076, p < 0.001, r_{rb} = 0.75$)。下位尺度についても同様に、Emotional Responsiveness (PETS-ER) において共感的対話条件 ($Mdn = 78.8$) は CTDS 条件 ($Mdn = 60.0$) よりも有意に高く ($W = 3992, p < 0.001, r_{rb} = 0.71$)、Understanding and Trust (PETS-UT) においても共感的対話条件 ($Mdn = 78.0$) が CTDS 条件 ($Mdn = 60.5$) を有意に上回っていた ($W = 4059, p < 0.001, r_{rb} = 0.74$)。

以上の結果より、共感的対話条件ではシステムが高い共感性を示し、一方で CTDS 条件では共感性が抑制され、意図した操作が参加者に適切に知覚されていたことが確認された。

5.3 対話戦略と好感度が認知的変化に及ぼす影響

ユーザのエージェントに対する好感度が + のとき、CTDS は共感的対話戦略と比較して、認知的変化を強く喚起するという仮説を検証するために、好感度の影響を調整した上での対話戦略本来の効果を調査した。具体的には、SERI Reappraisal スコアを目的変数、対話戦略 (共感/CTDS) と Godspeed Likeability スコア、および IOS Scale スコアを固定効果、参加者をランダム効果とした線形混合モデル (LMM) による分析を行った。

分析の結果、対話戦略の主効果に有意な差が認められ ($F(1, 114) = 5.11, p = 0.026$)、CTDS 条件の修正平均 ($M = 4.55, SE = 0.132$) は共感的対話条件 ($M = 4.15, SE = 0.132$) よりも優位に高いスコアを示した (図 3)。これは、エージェントへの好感度が同程度である場合、CTDS 条件の方が共感条件よりも有意に高い認知的変化をユーザに喚起することを示している。また、Likeability の主効果も有意であり ($F(1, 178) = 12.99, p < 0.001$)、エージェントへの好感度が高いほど、認知的変化を起こしやすい傾向が確認された。図 4 は、Likeability と Reappraisal の関係を示したもの

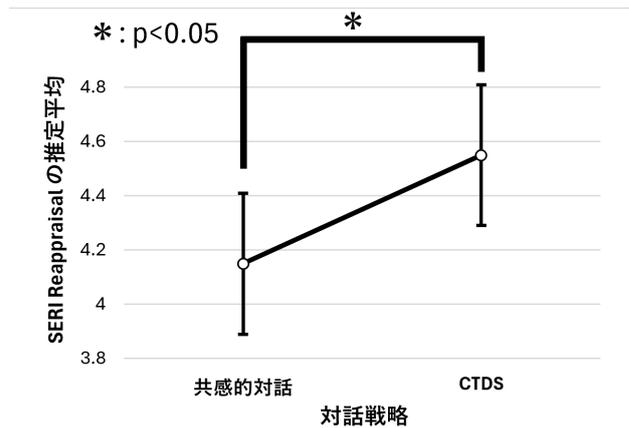


図 3: 対話戦略を主効果とした場合の SERI Reappraisal の推定周辺平均値（エラーバーは 95%CI）

である。高い Godspeed Likeability を示した参加者群（分布の右側に位置する層）の SERI Reappraisal スコアの分布を見ると、共感的対話条件が 2 点から 7 点まで満遍なく分布している一方で、CTDS 条件では 4 点以上に分布が偏っていることがわかる。これらの結果から仮説が支持された。

5.4 対話戦略が好感度に及ぼす影響

次に、CTDS 発話が好感度に与えた影響について検証するために、全データについてウィルコクソンの符号順位検定を行った。Godspeed Likeability スコアおよび IOS Scale スコアのいずれの指標においても CTDS 条件は共感条件と比較して有意に低い値を示した ($W = 2852, p < 0.001, r_{rb} = 0.72$ および $W = 2477, p < 0.01, r_{rb} = 0.74$)。Godspeed Likeability の分布 (図 4) を確認すると、共感条件では高得点帯 (3 点以上) に偏っているのに対し、CTDS 条件では分布が全体に広がっており、CTDS によって対話エージェントの好感度を大きく下げた参加者が一定数存在したことがわかる。

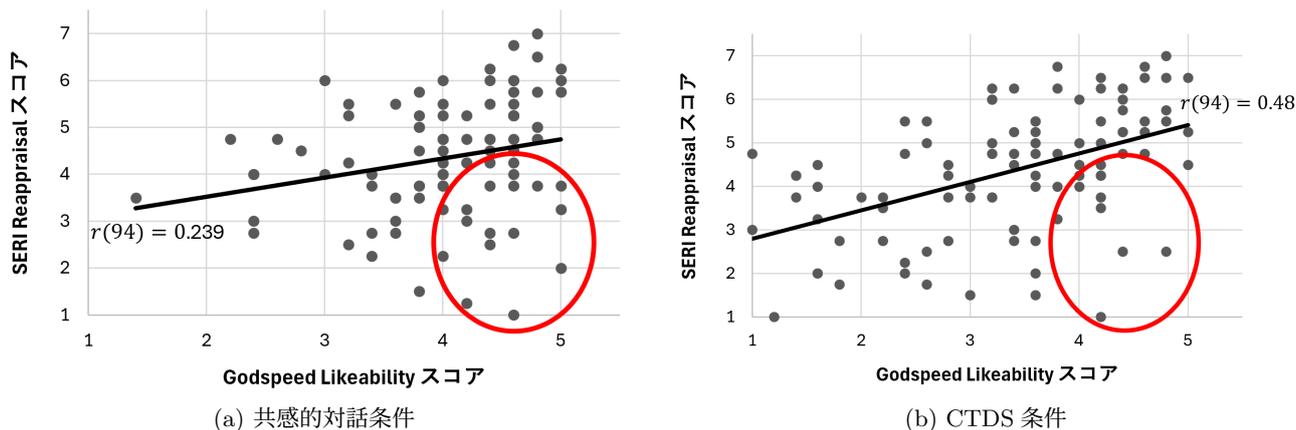


図 4: Godspeed Likeability と SERI Reappraisal のスコア分布

5.5 認知的変化の全体比較

SERI Reappraisal スコア全体の平均値をウィルコクソンの符号順位検定により比較した。その結果、両条件間に統計的な有意差は認められなかった ($W = 1768, p = 0.796, r_{rb} = -0.03$)。5.3 節と 5.4 節の結果を統合すると、これは CTDS 自体は共感的対話よりも高い認知的変化の促進効果を持つ (5.3 節) もの、多くの参加者において好感度の低下を招いた (5.4 節) ことにより、そのプラス効果が相殺されてしまった結果であると解釈できる。

5.6 対話戦略と好感度がユーザの不安解消に与える効果

次に、CTDS がユーザの不安解消に効果的かを検証するため、STAI A-state スコアを目的変数、対話戦略 (共感/CTDS) と Godspeed Likeability スコア、および IOS Scale スコアを固定効果、参加者をランダム効果とした線形混合モデル (LMM) による分析を行った。分析の結果、Godspeed Likeability スコアの主効果に有意な差が認められた ($F(1, 180) = 16.23, p < 0.001$) 一方で、対話戦略の主効果に有意な差は認められなかった ($F(1, 111) = 0.99, p = 0.322$)。

5.7 性格特性が CTDS の評価指標に与える影響

最後に、参加者の性格特性が CTDS 条件における認知的変化の喚起 (SERI Reappraisal)、エージェントへの好感度 (Godspeed Likeability, IOS Scale) にどのような影響を与えるかを検討するため、強制投入法による重回帰分析を実施した。

説明変数には、ERQ-J の再評価方略スコア、抑制方略スコア、および STAI の特性不安 (A-trait) スコアの 3 項目を投入した。従属変数は、SERI Reappraisal スコア、Godspeed Likeability スコア、IOS Scale スコアの 3 つを設定し、それぞれについて個別に分析を行った。なお、検定の多重性を考慮してボンフェローニ法による補正を行い、有意水準は $\alpha = 0.05/3 \approx 0.017$ とした。

分析の結果、SERI Reappraisal スコアおよび IOS Scale スコアを従属変数としたモデルでは、有意な関連は確認されなかった。一方、Godspeed Likeability スコアを従属変数としたモデルにおいては、決定係数 $R^2 = 0.084$ であった。各変数の影響を確認したところ、ERQ-J 再評価方略スコアにおいて有意な負の影響が認められた ($B = -0.35, SE = 0.12, t = -2.81, p = 0.006 < 0.017$)。これ以外の ERQ-J 抑制方略スコア ($p = 0.195$) および STAI 特性不安スコア ($p = 0.247$) については有意な影響は見られなかった。

この結果は、日常的に認知的変化を行う傾向が高い参加者ほど、CTDS エージェントに対して低い好感度を抱く傾向があることを示唆している。

6 議論

本研究では、インタラクションモデルに基づく CTDS 対話エージェントを実装し、実対話実験を通してその有効性を検証した。以下では、先行研究 [8] のビデオベース実験と本研究の対話実験結果から得られた知見について議論する。

6.1 実対話における CTDS の有効性とインタラクションモデルの妥当性

実験の結果、エージェントへの好感度の影響を統計的に調整した場合、CTDS 条件は共感的対話条件よりも有意に高い認知的変化を喚起することが示された (5.3 節)。これは、「好感度が維持されている」という前提において、CTDS がユーザの認知的変化を促すのに有効な戦略であること、すなわち本研究で提案したインタラクションモデルの成功ルート (図 1: 成功) が理論的に妥当であることを示唆している。一方で、好感度を考慮しない全体比較 (5.5 節) では認知的変化に有意差が見られず、かつ CTDS 条件の好感度は著しく低下していた (5.4 節)。これは、モデルにおける失敗ルート (図 1: 失敗)、すなわち「バランス回復のためにエージェントへの好感度を悪化させる (ユーザ → エージェントを - にする)」という現象が実対話において顕著に現れた結果と解釈できる。CTDS は本来、認知的変化を促すポテンシャルを持つが、好感度の低下がその効果を相殺してしまうというトレードオフの関係が明らかとなった。

6.2 実対話における好感度低下の要因

先行研究 [8] のビデオベース実験では CTDS による好感度の有意な低下は見られなかったが、本実験では有意な低下が確認された。この乖離の主な要因として、評価視点の違い (当事者性) が考えられる。ビデオ実験のような第三者視点では、CTDS 発話は「論理的なアドバイス」として客観的に評価されやすい。しかし、本実験のような第一人称視点の実対話では、自身の抱える不安やネガティブな感情を直接否定される形となるため、攻撃性や拒絶として知覚されやすく、結果として失敗ルート (好感度低下) へ遷移しやすかったと考えられる。また、本実験では「友人 (ツレ)」という役割設定はあったものの、実際には初対面のエージェントと 15 ターンという短期間の対話を行ったに過ぎない。バランス理論 [9] に基づけば、CTDS が機能するには、前提として自然に悩みを相談するようなポジティブ関係 (ユーザ \rightarrow エージェント) が必要である。十分なラポールが形成されていない段階での性急な CTDS の導入は、関係性の破綻を招くリスクが高い。したがって、実際の運用においては、まずは共感的対話で十分に好感と信頼関係を構築し、ある程度関係性が形成されたタイミングで CTDS に切り替えるといった、長期的かつハイブリッドな戦略が必要不可欠であると示唆される。

6.3 認知的変化の特性と CTDS

興味深い知見として、日常的に認知的再評価を行う傾向 (ERQ-J 再評価方略スコア) が高いユーザほど、CTDS エージェントに対して低い好感度を示す傾向が見られた (5.7 節)。本来、認知的変化が得意なユーザは CTDS の意図を理解しやすいとも考えられるが、結果は逆であった。これは、自力で認知的変化が可能ユーザにとって、エージェントからの批判的な指摘や視点を切り替えるよう働きかける提案は、「単に否定的なエージェント」もしくは「お節いなエージェント」としてネガティブに知覚された可能性を示唆している。逆に言えば、自力での認知的変化が苦手なユーザに対してこそ、CTDS は外部からの介入として有効に機能する可能性がある。このことから、CTDS の実装にあたっては、ユーザの特性やその場の感情処理能力に応じて、介入の有無や強度を調整するパーソナライゼーションが重要であると考えられる。

6.4 CTDS がユーザとの関係性深化に与える効果

本節では、ソシオン理論 [18] に基づき、CTDS エージェントがユーザと長期的なインタラクションを重ねることで構築される関係性について議論する。

ソシオン理論 [18] において、他者との関係性の深さは、個体間を結ぶコミュニケーションチャンネルの荷重の大きさとして定義される。この荷重は信頼や情報の信憑性 (リアリティ) として機能し、その値は個々のソシオンの学習プロセスを通じて決定され、動的に更新されるものである [18]。荷重の増大に寄与する主要な要因として、互酬性の規範および経験的学習が挙げられる。互酬性の規範とは、「自分と他者の荷重の対象性を強める」という規範であり、相手が自分を信頼してくれている (自分に高い荷重を置いている) と知れば、自分も相手への信頼 (荷重) を高めようとする傾向である [18]。経験的学習 (成功と失敗) とは「信じてうまくいったら、ますます信じる」というルールであり、相手の情報を信頼して良い結果が得られれば荷重を強め、逆に「痛い目」にあえば荷重を撤収 (関係を希薄化) するものである [18]。

これらのメカニズムを CTDS に適用すれば、適切なインタラクションを通じて、長期的にはユーザとの関係性を深化させることが可能であると考えられる。具体的には、第一に、CTDS がユーザの好感度を損なうことなく認知的変化を喚起することに成功した場合、経験的学習の効果により、ユーザはエージェントに対する荷重 (信頼) を強化する。第二に、CTDS による一見批判的に見える言動が、実はユーザを鼓舞するための意図的な発話 (エージェントがユーザに高い荷重を置

いていることの出)であるとユーザが認識した場合、互酬性の規範に基づき、ユーザ側もエージェントへの荷重を高めることで関係性が深化すると考えられる。

近年の研究では、エージェントを単なるツールとしてではなく、チームメイトやパートナーとして位置付ける試みが行われている [19, 20]。本研究で示唆された関係性深化の可能性は、こうした人間とエージェントの共生関係を目指す研究アプローチを、より一層発展させるための重要な知見となると考えられる。

7 おわりに

本研究では、ユーザのネガティブな感情に対し敢えて非同意な立場をとることで認知的変化を促す Critical Tongue Dialogue Strategy (CTDS)[8] の実対話での効果検証を行った。具体的には、新たに設計したインタラクティブモデル駆動プロンプト (IMDP) を用いて対話エージェントに実装し、対話実験を行った。実験の結果、CTDS は共感的対話と比較して、統計的に有意に強く認知的変化を喚起することが示された。この結果は CTDS のインタラクティブモデルが、ビデオベースの受動的な評価だけでなく、実際の対話においても機能し、ユーザの認知的変化を促進する有効な手段となり得ることを示している。

一方で、CTDS のインタラクティブモデルで予測されていたとおり、実対話においてはエージェントに対する好感度の低下が確認された。これは、CTDS の発話が第三者視点では「有益な助言」として受け取られた一方で、当事者視点では攻撃的、否定的な発話として知覚されたことにより生じた可能性がある。この結果は CTDS のインタラクティブモデルが実際の感情変容の傾向を反映していることを裏付ける結果であるが、同時に好感度の低下は CTDS にとって重大な副作用の一つである。

今後の展望としては、CTDS の好感度低下を抑制する手法の開発が必要である。具体的には、CTDS 発話の受容性を高めるための穏やかな言い回しの発話生成手法の開発や、効果的なフォローアップ発話の追加などが考えられる。加えて本研究の対話実験結果から示唆されたように、ユーザの性格特性やその時の心理状態に応じて、CTDS の適用可否やタイミングを動的に調整するパーソナライゼーション機構の開発も課題となる。

謝辞

本研究は、JSPS 科研費 (JP23K16923) の支援を一部受けました。記して感謝いたします。

参考文献

- [1] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*, Vol. 6, No. 11, p. e12106, Nov 2018.
- [2] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res*, Vol. 20, No. 6, p. e10148, Jun 2018.
- [3] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. Embodied conversational agents in clinical psychology: A scoping review. *J Med Internet Res*, Vol. 19, No. 5, p. e151, May 2017.
- [4] R. Friedman, C. Anderson, J. Brett, M. Olekalns, N. Goates, and C. C. Lisco. The positive and negative effects of anger on dispute resolution: Evidence from electronically mediated disputes. *Journal of Applied Psychology*, Vol. 89, No. 2, pp. 369–376, 2004.
- [5] Hannah Gaffney, Warren Mansell, and Sara Tai. Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Ment Health*, Vol. 6, No. 10, p. e14166, Oct 2019.
- [6] Wojciech Marek Kulesza, Aleksandra Cislak, Robin R. Vallacher, Andrzej Nowak, Martyna Czekiel, and Sylwia Bedynska and. The face of the chameleon: The experience of facial mimicry for the mimicker and the mimickee. *The Journal of Social Psychology*, Vol. 155, No. 6, pp. 590–604, 2015. PMID: 25811746.
- [7] Adam Miner, Amanda Chow, Sarah Adler, Ilia Zaitsev, Paul Tero, Alison Darcy, and Andreas Paepcke. Conversational agents and mental health: Theory-informed assessment of language and affect. In *Proceedings of the Fourth International Conference on Human Agent Interaction, HAI '16*, p. 123–130, New York, NY, USA, 2016. Association for Computing Machinery.

- [8] Keisuke Magara, Tomoki Miyamoto, and Akira Utsumi. The role of a critical tongue dialogue strategy in stimulating emotion regulation: An interaction model and video-based study. In *Proceedings of the 13th International Conference on Human-Agent Interaction, HAI '25*, p. 10–18. Association for Computing Machinery, 2026.
- [9] Fritz Heider. *The Psychology of Interpersonal Relations*. Lawrence Erlbaum Associates, New York, 1958.
- [10] 眞柄圭佑, 宮本友樹, 内海彰. ユーザの不安軽減を目的とする対話エージェントの毒舌を交えた非同意的対話戦略の検討. In *Human-Agent Interaction Symposium 2025*, Osaka, Japan, 2025.
- [11] James J Gross. Emotion regulation: affective, cognitive, and social consequences. *Psychophysiology*, Vol. 39, No. 3, pp. 281–91, May 2002.
- [12] Jun Yoshizu, Rikuko Sekiguchi, and Toshihiko Amemiya. Development of a japanese version of emotion regulation questionnaire. *JAPANESE JOURNAL OF RESEARCH ON EMOTIONS*, Vol. 20, No. 2, pp. 56–62, 2013.
- [13] Benjamin A Katz, Neta Lustig, Yael Assis, and Iftah Yovel. Measuring regulation in the here and now: The development and validation of the state emotion regulation inventory (seri). *Psychological assessment*, Vol. 29, No. 10, p. 1235, 2017.
- [14] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. Perceived empathy of technology scale (pets): Measuring empathy of systems toward the user. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, New York, NY, USA, 2024. Association for Computing Machinery.
- [15] 清水秀美, 今栄国晴. State-trait anxiety inventory の日本語版 (大学生用) の作成. *The Japanese Journal of Educational Psychology*, Vol. 29, No. 4, pp. 348–353, 1981.
- [16] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, Vol. 1, No. 1, pp. 71–81, jan 2009.
- [17] Arthur Aron, Elaine N. Aron, and Danny Smolman. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, Vol. 63, No. 4, pp. 596–612, 1992.
- [18] 木村洋二, 藤沢等, 雨宮俊彦. ソシオンの理論: ソーシャル・ネットワークへのシステム・ダイナミック・アプローチ. 関西大学社会学部紀要, Vol. 21, No. 2, pp. 67–143, 1990.
- [19] Shogo Kanda, Masayuki Kanbara, Taishi Sawabe, Yuichiro Fujimoto, and Hirokazu Kato. Robot to play video games together. In *Proceedings of the 11th International Conference on Human-Agent Interaction, HAI '23*, p. 238–245, New York, NY, USA, 2023. Association for Computing Machinery.
- [20] Isabella Seeber, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, and Matthias Söllner. Machines as teammates: A research agenda on ai in team collaboration. *Information & Management*, Vol. 57, No. 2, p. 103174, 2020.