# LLM を利用したテキスト対話エージェントの返信間隔がユーザに与える印象の調査

## An Investigation How the Reply Intervals of an LLM-Based Text Dialogue Agent Affect Users' Impressions

神保一馬 [1]　小松孝徳 [2]

Kazuma Jimbo[1] and Takanori Komatsu[2]

[1] 明治大学院先端数理科学研究科，[2] 明治大学総合数理学部
[1] Graduate School of Advanced Mathematical Sciences, Meiji University
[2] School of Interdisciplinary Mathematical Sciences, Meiji University

雑談を目的としたテキスト対話エージェントの返信間隔とそのエージェントへの印象を調査した先行研究では，1 時間という比較的長い返信間隔のエージェントがユーザから高く評価されることが明らかとなった．しかし，その先行研究で使用されたエージェントは参加者の発話内容に関係なく，あらかじめ決められた内容の返信をするものであった．そこで本研究では，LLM を利用し参加者の発話内容を踏まえた自然な返信を生成するエージェントを作成し，先行研究同様に返信間隔とそのエージェントへの印象を調査した．その結果，LLM を利用したエージェントは，先行研究のエージェントが苦手としていた 1 分という返信間隔において評価を向上させていることが確認されたが，それ以外の返信間隔では特に評価を向上させてはいないことが確認された．このことから，返信間隔に適した返信内容を設計することができれば，LLM を使用しない固定の返信を行うエージェントでも高い評価を得ることができると明らかになった．

## 1. はじめに

昨今の急速な AI 技術の発展により，ユーザとテキスト対話を行うテキスト対話エージェントが普及しつつあり，それらはチャットボットとして様々なシステムに実装されている．例えばヤマト運輸は，ユーザから送られるテキストメッセージによって配達時間や配達方法を指定することができる LINE bot を開発しており[1]，任天堂ではユーザからの簡単な問い合わせに対応するヘルプデスクとしてチャットボットを活用している[2]．チャットボットの多くは，上記のように問い合わせ対応などの具体的な業務を担当している一方，人間と自然な雑談を行うことを目的としたテキスト対話エージェントも開発されつつある．

その一例として挙げられるのが，Microsoft 社によって開発された AI コミュニケーション bot の「りんな」である[3]．「りんな」はディープラーニングを用いた自然言語処理技術により，ユーザと人間のような文脈を踏まえた自然な会話を行うことができる．しかしながら，「りんな」との会話においてユーザが「死ね」という単語を多用するなど，必ずしも「りんな」はユーザに受け入れられていなかったと Microsoft の研究チームは報告した．そこで著者らは，「りんな」がユーザから受けられなかった原因の一つとして，ユーザからのメッセージを受け取ると即座に返信を行うという即時応答性にあると考えた．人間同士のテキストチャットでは，メッセージの送信後，その返信までの間隔は数秒間空くことは普通であり，返信が即座に返ってくることはほとんどない．そのため，「りんな」の即座の返信がユーザに対して違和感や圧迫感を与えていたと考えられた．

この問題に対して著者らは，人間とエージェントのテキストを用いたコミュニケーションの円滑化を目指して，ユーザとエージェントの 1 対 1 での対話において返信間隔の違いがユーザからの印象に与える影響を調べる研究を行った[4]．その結果，Moon らの研究[5]で明らかになっていた 5 秒という短い間隔の他に，1 時間という適度に長い間隔で返信を行うエージェントも参加者から高い評価を得られることが明らかになった．しかし，この研究で使用さ

れたエージェントは参加者の発話内容に関係なく，あらかじめ決められた内容の返信を行っていた．そのため，返信間隔とエージェントの評価について得られた結果が，実験における対話の内容に依存していた可能性があった．

そこで本研究では，LLM を利用して自然な返信を行うテキスト対話エージェントが上記の研究[4]と同様の対話を行った際に，返信間隔がエージェントからの印象に与える影響を調査した．具体的には，ユーザによるメッセージの送信からエージェントによる返信までの間隔（交替潜時）を「独立変数」，その際のユーザのエージェントへの印象を「従属変数」としたインタラクション実験を実施した．

## 2. 関連研究

著者らは，人間とテキスト対話エージェントによる 1 対 1 での対話に適した設計を調べるために，返信間隔の違いがユーザからの印象に与える影響を調べる研究を行った[4]．返信間隔として 0 秒，5 秒，1 分，10 分，1 時間，24 時間の 6 水準を設定した．その結果，先行研究[5]で明らかになっていた 5 秒という短い間隔の他に，1 時間という適度に長い間隔で返信を行うエージェントも参加者から高い評価を得られることが明らかになった．5 秒という返信間隔で高い評価が得られた要因は，即座の返信ではないためエージェントが考えて返信を行っている印象を参加者に与えられたこと，返信までの時間がシステムとして許容できる程度の長さであったことだと考えられた. また, 1 時間という返信間隔で高い評価が得られたことについて， LINE などの SNS で人間同士が実際に対話を行う際の返信間隔に近い間隔であると考えられた．しかしこの研究では，エージェントからの返信内容はあらかじめ決められており，参加者の発話に応じた返信を行う機能は有していなかった．そのため，一部の参加者は，自由記述にて「質問を無視された」「会話が噛み合わなかった」と回答していた．そのため，参加者からのメッセージに応じた自然な返信を行うテキスト対話エージェントにおいても，同様の返答間隔で高い評価が得られるかどうかは明らかになっていないといえる．

Moon [5]は，参加者に返信間隔が操作されたテキスト対話エージェントと共に「砂漠の生存問題」に取り組んでもらう実験を実施した.「砂漠の生存問題」とは, 砂漠で生き延びるために必要な道具 12 個の中から，どの道具を使用したいかのランク付けを行うというものである．その結果，短い返信間隔（0〜1 秒）や長い返信間隔（13〜18 秒）よりも，適度な返信間隔（5〜10 秒）のエージェントの方が，ユーザにとって説得力があると認識されていたことが明らかになった．

実際の人間同士の対話において自分が知らない話題について話す際には，相手に質問をしたり，相手から得た情報についての自分の感想を述べたりすることが多い．一方で，自分が良く知っている話題について話をするときには，自分の知識に基づいた内容の発話を行うことが多い．加藤ら[6]はこのことに着目して，非タスク志向の音声対話において，話題に対する親密度に応じて異なる応答を行う対話システムを提案した．ここで，親密度とは話題に関する知識の豊富さを意味する．実験においては,「うん」「なるほど」といったシンプルな応答を行うシステム，詳細を尋ねる質問や自分の感想といった親密度の低い応答を行うシステム，同意や共感を示す応答を頻繁に行う親密度の高いシステムの合計 3 つのシステムをランダムな順番で実験参加者に使用してもらい，アンケートを用いて主観評価を行なってもらった. その結果, 親密度の低い応答をするシステム，親密度の高い応答をするシステム，両方の有効性が示された．親密度の低い応答をするシステムは，聞き役として良い印象をユーザに与えられること，比較的良い満足度をユーザに与えられることを確認した．親密度の高い応答をするシステムは，ユーザの話に対して関心を持っていると感じさせられること，システムとまた話してみたいとユーザに感じさせられることを確認した．ただし，親密度の高い応答をすることで，聞き役ではなく，話し役としての役割が強くなる可能性も示唆された．このことから，エージェントが聞き役である場合と話し役である場合には異なる評価が行われることが明らかになった．

先行研究におけるエージェントは固定の返信を行うため常に聞き役であったのに対して，本研究のエージェントは内容に応じた返信を行うため聞き役と話し役の両方を行うことができる．そのため，先行研究と本研究ではエージェントの評価の傾向に違いが生じる可能性が高く，LLM を利用した返信を行うエージェントについて返信間隔を変化させた際の評価の推移について調べることに意義があるといえる．

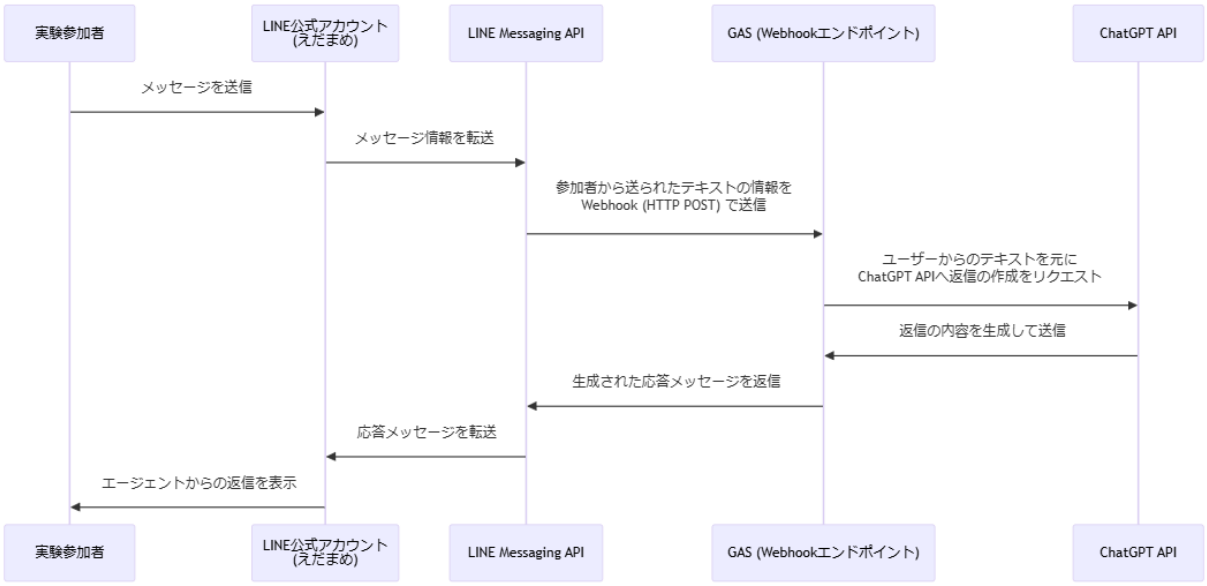


図 1：「えだまめ」の構成

# 3. 実験

## 3. 1. テキスト対話エージェント「えだまめ」

本実験にて参加者とテキスト対話を行うエージェントとして，LINE bot「えだまめ」を作成した．「えだまめ」は LINE Developer および Google App Script にて実装された bot であり，LINE 上での公式アカウントとして運用した（図 1）．

「えだまめ」は実験参加者からメッセージを受け取ると，実験条件に対応した返信間隔でメッセージを返信する．具体的に「えだまめ」は，ChatGPT API の gpt-4o-2024-11-20[1]モデルを利用して，参加者のメッセージの内容に応じて，自然な返信を行うように設計された．対話の内容ではなく返信間隔の違いが参加者の心象に与える影響について調査するためには，参加者の送信内容によって「えだまめ」と参加者との対話内容が大きく変化することを防ぐ必要がある．そこで，temperature を 0.1 と低く設定し，モデルが生成する話題についても「趣味」「最近楽しかったこと」に限定した．temperature は，LLM の出力を決定する確率分布の散らばり具合を決定する値であり，値が 0 に近づくほど応答のランダム性が低下する. そのため, temperature を低く設定することで，自然な応答をしながら参加者間の対話内容の差を小さくできる．

## 3. 2. 実験設定

本実験では，ユーザのメッセージ送信から「えだまめ」の返信までの間隔を独立変数として設定した．独立変数の具体的な水準として，0 秒，5 秒，1 分，10 分，1 時間，8 時間の 6 水準を設定した．先行研究と比較を行うために，先行研究と同一の返信間隔を設定した一方で, 先行研究における 24 時間水準を 8 時間水準に変更した．先行研究における 24 時間水準は「返信までの間隔があまりに長く，実験が正常に進行しているか不安になった.」といった記述が複数みられた．このことから，24 時間という返信間隔で実験を行った際に，返信間隔以外に実験設計への不安がエージェントや対話の評価に影響してしまっていることが明らかになった．したがって，24 時間という返信間隔で評価を適切に調査することは難しいと考え，水準から除外した．しかし，返信間隔を変化させた際のエージェントの印象評価の推移について調べるためには，先行研究で評価のピークとなっていた 1 時間という水準よりも長い間隔での調査を行う必要がある. 以上を踏まえて, 1 時間よりも長く 24 時間よりも短い水準として 8 時間という水準を新たに設定した．

「えだまめ」に対するアンケート調査の結果を従属変数として設定した（表 1）．比較を行うために，アンケートは先行研究と同一のものを用いた．本アンケートは, b-1 から b-10 までの 5 段階の 10 問のリッカート尺度（最低評価が 1，最高評価が 5：b-5 は

[1] https://platform.openai.com/docs/models/#gpt-4o

表 1：実験後アンケートの質問項目

| | 質問 |
|---|---|
| b-1 | 「えだまめ」は親しみやすいと感じた |
| b-2 | 「えだまめ」と仲良くなれたと感じた |
| b-3 | 「えだまめ」には人間味があると感じた |
| b-4 | 「えだまめ」は好印象だった |
| b-5 | 「えだまめ」との会話に圧迫感を感じた |
| b-6 | 「えだまめ」は話しかけやすいと感じた |
| b-7 | 「えだまめ」との会話は返信しやすいと感じた |
| b-8 | 「えだまめ」との会話は盛り上がったと感じた |
| b-9 | 「えだまめ」と話すのは楽しいと感じた |
| b-10 | 「えだまめ」との会話は自然だと感じた |

逆転項目）から構成されている．本実験ではこの 10 問の質問項目を，「えだまめ」というエージェントの人格に対する評価である「人格的印象点」と，「えだまめ」との会話自体に対する評価である「機能的印象点」の二種類に分類して扱うこととした．具体的には，b-1 から b-4 までの四問の合計点を「人格的印象点」（最低点 4 点，最高点 20 点，クロンバックの $\alpha$ 係数：0.88），b-5 から b-10 までの 6 問の合計点を「機能的印象点」（最低点 6 点，最高点 30 点，クロンバックの $\alpha$ 係数：0.87）とした．表 1 の質問のほかに，「えだまめ」に対する印象を自由に記入してもらう自由記述欄にも回答を求めた．

また，「えだまめ」との対話における参加者の発話文字数も従属変数として設定した．発話文字数は「えだまめ」との対話で参加者が行ったすべての発話の文字数の合計である．

## 3. 3. 実験参加者

本実験は参加者間計画として実施され，63 名の大学生および大学院生が参加した．これらの参加者は講義の課題の一環として本実験に参加したため，報酬は支払われなかった．また，これら 63 名の参加者は無作為に独立変数の 6 水準に配置された．各水準に配置された人数は 0 秒水準：12 人，5 秒水準：11 人，1 分水準：10 人，10 分水準：11 人，1 時間水準：11 人，8 時間水準：8 人となった．なお，参加者にはテキスト対話エージェントの返信間隔については何も情報を与えなかった．

## 3. 4. 実験手順

参加者はまず，配置された実験条件に応じた返信間隔が設定された「えだまめ」を LINE で友達に追加するように指示された．「えだまめ」を友達に追加した後，参加者が「えだまめ」に話しかけることで，

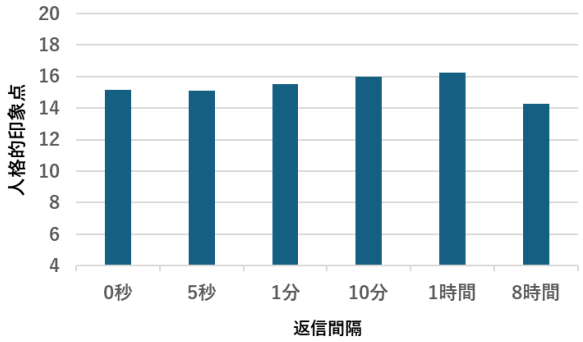


図 2：LLM を使用したエージェントの人格的印象点

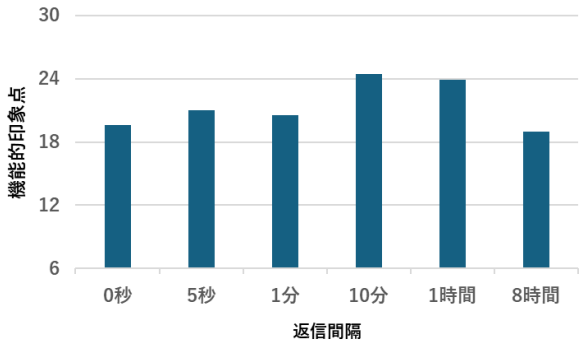


図 3：LLM を使用したエージェントの機能的印象点

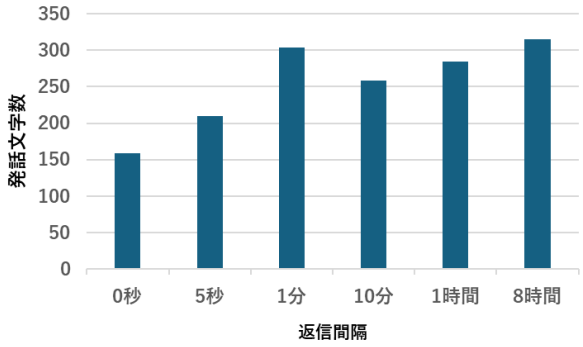


図 4：LLM を使用したエージェントとの対話における参加者の発話文字数

「えだまめ」との対話が開始された．先行研究と同様に「えだまめ」がメッセージを 14 回送信した段階で会話を終了とした．そして会話終了時に「えだまめ」から示される URL にアクセスし，Google Form に実装された実験後アンケートに回答することで実験が終了した．

また，「えだまめ」との会話の途中で 7 日間以上返信をしなかった参加者は会話を放棄したとみなし，

会話が終了していなくても実験後アンケートに回答してもらい，同時に会話を放棄した理由を記載す

るように指示された．

### 3. 5. 実験結果

人格的印象点に関して，一要因参加者間分散分析（独立変数：返信間隔（6 水準），従属変数：人格的印象点）を行ったところ，群の効果は有意でなかった[F(5,62) = 0.27, n.s.]（図 2）．続いて，機能的印象点に関して，一要因参加者間分散分析（独立変数：返信間隔（6 水準），従属変数：機能的印象点）を行ったところ，群の効果は有意でなかった [F(5,62) = 1.56, n.s.]（図 3）．対話における参加者の発話の文字数合計の平均値について一要因参加者間分散分析（独立変数：返信間隔（6 水準），従属変数：発話文字数）を行ったところ，群の効果に有意傾向が観察されたものの [F(5,62) = 2.01, p <. 10]（図 4），HSD 法による多重比較の結果，どの水準間においても有意差は観察されなかった（MSe = 19268.53, p > .05）．

そこで，本研究との結果と先行研究の結果と比較するために，返答間隔水準ごとに人格的印象点について Welch の t 検定を行った．その結果，1 分水準において有意差が観察され（p < .05），LLM を使用した返信を行うエージェントの評価が，固定の返信を行うエージェントよりも有意に高いことが明らかになった（図 5）．同様に機能的印象点について Welch の t 検定を行ったところ，0 秒水準において有意差が観察され（p < .05），LLM を使用した返信を行うエージェントの評価が，固定の返信を行うエージェントよりも有意に低いことが明らかになった（図 6）．また，10 秒水準において有意傾向が見られ（p < .10），LLM を使用した返信を行うエージェントの評価が，固定の返信を行うエージェントよりも高い傾向が観察された．同様に，返信文字数について Welch の t 検定を行ったところ， 1 分水準において有意傾向が見られ（p < .10），LLM を使用した返信を行うエージェントとの対話における発話文字数が，固定の返信を行うエージェントよりも文字数が多い傾向が観察された（図 7）．

なお，「えだまめ」との会話を放棄した参加者数は 0 秒水準: 1 人，5 秒水準: 4 人，1 分水準: 0 人，10 分水準: 1 人，1 時間水準: 2 人，8 時間水準: 3 人，であった．

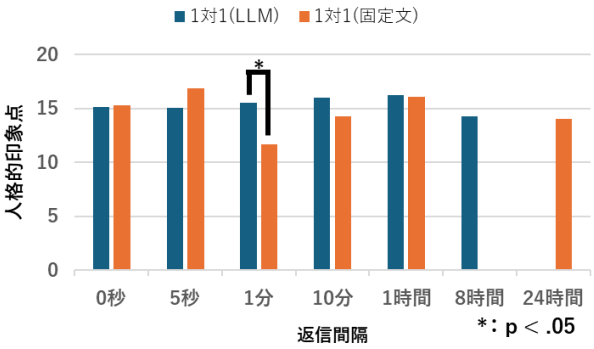


図 5：固定の返信を行うエージェントと LLM を使用した返信を行うエージェントの人格的印象点の比較

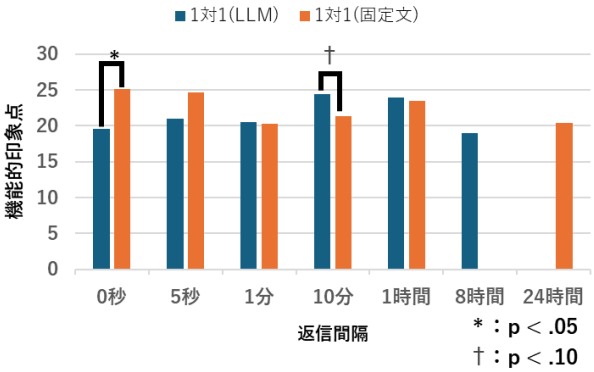


図 6：固定の返信を行うエージェントと LLM を使用した返信を行うエージェントの機能的印象点の比較

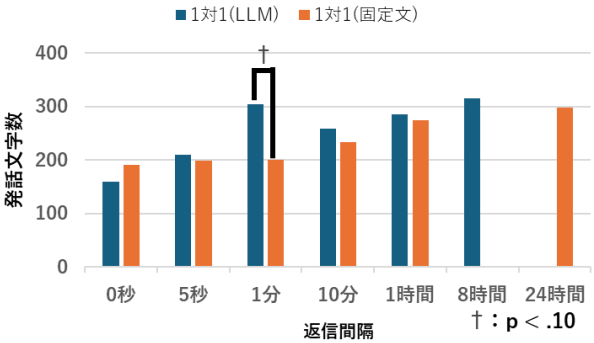


図 7：固定の返信を行うエージェントと LLM を使用した返信を行うエージェントの対話における参加者の発話文字数の比較

# 4. 考察

## 4. 1. 各水準における人格的印象点

本実験の結果より，エージェントに対する評価である人格的印象点は，1 時間水準において最も高い得点を示していることが分かったが，返信間隔間でエージェントの評価に統計的な有意差は観察されず，最大値である 1 時間水準の得点 16.27 と最小値である 8 時間水準の得点 14.25 の差が非常に小さいことから，水準間に差はなかったといえる（図 2）．一方，固定の返信を行うエージェントを使用した先行研究では，返信間隔が 5 秒と 1 時間の際にエージェントの評価が有意に高くなる 2 つのピークを持つ山型の分布となっていた（図 5）．ここから， LLM を使用して自然な返信を行うエージェントとの対話においては，返信間隔はエージェントの評価に影響を与えていなかったという，固定の返信を行うエージェントを用いた先行研究は異なる結果が観察されたといえる．

また LLM を使用したエージェントは先行研究において評価の低かった 1 分水準，10 分水準において評価を向上させていた一方で，先行研究で高い評価を得ていた 0 秒，5 秒，1 時間水準では同程度の評価を得ていたことが観察された．特に，0 秒，5 秒水準においては本研究よりも先行研究におけるエージェントの評価が高かったといえる（図 4）．

その原因について，著者らは以下のように考えている．固定の返信を行う先行研究のエージェントは，参加者の返信に対して内容に関わらず意味が通じるような当り障りのない相槌や質問を行うように設計されていた一方，本研究のエージェントは返信の内容に応じた具体的な内容を含む返信を行っていた．実際に，本研究の対話後の自由記述において，どの返信間隔水準においても「特定の人やモノの話題にも的確な返答が行われたことに驚いた」という感想が複数観察された．このことから，本研究のエージェントは LLM を利用することで自然な返信が可能になり，対話が弾んだことで参加者はより長い返信を行うようになったと考えられる．実際に参加者の発話文字数を確認すると，人格的印象点が有意に向上した 1 分水準においては，参加者の発話文字数も多いという傾向が観察された （図 7). したがって，参加者がより長い返信を行うようになったことでメッセージの内容について考える時間や入力する時間が長くなり，結果として 1 分や 10 分といった比較的長い返信間隔において，待つことによって感じるストレスが軽減されたことによって評価が向上したと推測された．先行研究においてみられた「1 分水準は返信を待つには長く，長期的な対話を行うには間隔が短いことから待つ時間がストレスとなった.」という記述が本研究ではみられなかったことも，この推測を裏付けているといえよう．

以上のことから，1 分水準において人格的印象点が有意に向上した要因は，LLM を利用したエージェントが参加者の発話内容に応じた返信を行うことで，返信間隔に適した対話を行ったことだと考えられた．逆に，5 秒水準では固定の返信を行うエージェントが，LLM を利用して自然な返信を行うエージェントの評価を上回っていたことから，先行研究のエージェントは 0 秒や 5 秒といった間隔で返信を行うのに適した内容の発話を行っていたと考えられた．したがって，返信間隔に適した返信内容を設計することができれば，LLM を使用しない固定の返信を行うエージェントでも高い評価を得ることができるといえる．

## 4. 2. 各水準における機能的印象点

本実験の結果より，対話内容に対する評価である機能的印象点は 10 分水準において最も高い得点を示していることが観察されたものの，返信間隔間で有意差は観察されなかった．つまり，LLM を使用して自然な返信を行うエージェントにおいては，返信間隔は対話内容の評価に影響を与えなかったことが確認された（図 3）．機能的印象点について，水準間の有意差が見られないという結果は先行研究と一致しているといえる（図 6）．

しかし，返信間隔ごとの機能的印象点の傾向には先行研究と本研究との間に差があることが確認された．例えば，0 秒，5 秒水準といった短い返信間隔においては先行研究のエージェントの評価の方が高く，10 分，1 時間水準といった長い返信間隔においては本研究のエージェントの評価の方が高いことが観察された（図 6）．つまり，短い返信間隔においては，対話内容に対する評価に関しても固定の返信が LMM を利用した返信よりも評価が高くなった．このことから，先行研究におけるエージェントとの対話における返信は短い返信間隔に適した内容であったと考えられ，これは 4.1 の考察と一致する．また，「参加者の返信の内容に関わらず対話が成り立つように設計したため，返信にはある程度の違和感がある．しかし，短い返信間隔の対話はテンポが速いため違和感に気が付きにくい」と先行研究では考察したが，本研究の実施によってこの考察が支持されたといえよう．

先行研究のように固定の返信を行うエージェントは運用コストが低いが，長い返信間隔の対話において違和感を抱かれやすい可能性がある．一方で，LLM を使用したエージェントは短い返信間隔においては必ずしも固定の返信よりも高い評価が得られるとはいえず，運用コストも高いのが実情である．よって，エージェントの使用目的に応じて，返信の生成方法を適切に選択することが重要であるといえる．

### 4. 3. 対話における参加者の発話文字数

先行研究では 24 時間水準において発話文字数が最も多かったことから,「対話の総時間が長くなるほどエージェントとのインタラクション時間が増加し，それに伴って発話文字数も増加する」と考察された．しかし本研究においては，対話における参加者の発話文字数は 8 時間水準で最も多いものの，1 分水準における文字数が 10 分，1 時間水準よりも多いことが明らかとなった（図 4）．このことから，返信間隔が 1 分を超える場合には，返信間隔はユーザの発話文字数に影響をさほど与えない可能性が示唆された．これは，ユーザがエージェントと関わりを持つのは返信を入力している間だけであるため，返信間隔が一定以上に長くなってもインタラクション時間自体は長くなっていないことに起因していたと推測できる．

本研究において 1 分水準で最も多い発話文字数が観察されたことから，メッセージを読み，返信を考え，入力するという一連の行為には 1 分程度あれば十分であり，それ以上の間隔を設定してもインタラクション時間は大きく増加しないと考えられた (図 4)．また，1 分水準における発話文字数が，返信間隔が長い 10 分水準や 1 時間水準よりも多くなった理由としては，人格的印象点が 1 分水準で最も高くなっていたことが考えられる(図 2, 図 4). したがって，発話文字数を増加させるためには返信間隔は 1 分程度で十分であると考えられた．

## 5. おわりに

本研究では, LLM を利用して自然な返信を行うテキスト対話エージェントがユーザと 1 対 1 で非タスク志向の対話を行った際に，返信間隔がエージェントの印象に与える影響を調査した. LLM を利用したエージェントは，先行研究のエージェントが苦手としていた 1 分という返信間隔において評価を向上させていることが確認されたが，それ以外の返信間隔では特に評価を向上させてはいないことが確認された．このことから，返信間隔に適した返信内容を設計することができれば, LLM を使用しない固定の返信を行うエージェントでも高い評価を得ることができると明らかになった．対話内容についても，短い返信間隔においては固定の返信を行うエージェントの評価が高くなった．また，ユーザの発話文字数を増加させるためには，返信間隔は 1 分程度で十分であることが示唆された．したがって，テキスト対話エージェントの設計においては，目的に応じた適切な返信間隔，返信の生成手法を選択することが重要であるといえる．

## 参考文献


[1] ヤマト運輸, “LINE で受け取るをもっと便利に”, https://www.kuronekoyamato.co.jp/ytc/campaign/renkei/LINE/, 参照 2024 年 1 月 23 日

[2] 任天堂, ”ソーシャルメディアアカウント”, https://www.nintendo.co.jp/social/, 参照 2024 年 1 月 23 日

[3] Xianchao Wu, Kazushige Ito, Katsuya Iida, Kazuna Tsuboi, Momo Klyen: “りんな：女子高生人工知能”, 言語処理学会 2016 年次大会 発表論文集, (2016)

[4] 須田翔悟, 神保一馬, 小松孝徳, 山田誠二 (2022).テキスト対話エージェントからの返信間隔の違いはユーザの印象評価に影響を与えるのか. HAI シンポジウム 2022.

[5] Moon. Y, “The effects of physical distance and response latency on persuasion in computer- mediated communication and human-computer communication”, Journal of Experimental Psychology: Applied, Vol.5, No.4, pp.379–392, (1999).

[6] 加藤大地, 原直, 阿部匡伸 (2020). 話題の対象に対する親密度に応じて応答する音声対話システムの検討. 研究報告音楽情報科学 (MUS), 2020(21), 1-6.