画像の言語化プロセスにおけるエージェントとのインタラクション形式の影響

Image Generation Method based on Subjective Impression Utterance

松岡竜輝 [1*]　今井倫太 [1]

[1] 慶應義塾大学
[1] Keio University

Abstract: 人が頭の中で思い描くイメージは，非常に曖昧なものである．そのため，そのイメージを言語化する際には，主観的な表現を用いる必要が生じる．一方で，画像生成モデルとのやり取りでは，モデルに対して客観的かつ明確な表現を意識する必要があり，この過程で言語化の負担が増大する．本研究では，エージェントを用いた画像生成の際に，エージェントへの入力が音声であるかテキストであるかで，人が入力する説明文にどのような違いが生じるかを調査し，メンタルワークロードへの影響を評価した．

1　はじめに

大規模言語モデル (LLM) の登場により，人とコンピュータのインタラクションでは，人が入力可能な言語表現が大幅に増加し，入力の自由度が大幅に増加した．一方，人がエージェントを意図した通りにコントロールするためには，エージェントに「最適」な命令文（プロンプト）を与える必要がある．しかしながら，人にとって，エージェントに対して最適なプロンプトを入力することは難しく，プロンプトが最適なものになるように試行錯誤を繰り返す必要がある．この問題の原因は，人-エージェントのインタラクションが指示に限定されていることにある．人のコミュニケーションには，指示を伝える「指示型コミュニケーション」と，お互いに相手の心の状態を推察し合う「共感型コミュニケーション」がある．モノの運搬依頼等で発生する指示型のコミュニケーションでは，客観的な言葉のみで意図の伝達が可能である．一方，イラスト作成などの創造的なタスクにおけるコミュニケーションは，依頼者が作成者に対して，依頼者の感覚に基づいたイメージを言語化する必要があり，主観的な言葉が増える．主観的な言葉は，解釈に対する個人差が客観的な言葉に比べて大きいため，自然と相手の心の状態（目的，信念，欲求，感情）や理解度を推察する「共感型のコミュニケーション」となる．人-エージェントのインタラクションでも，創造的なタスクにおけるインタラクションは共感型である必要があるが，従来の AI エージェントは道具として設計されており，人が客観的な指示を与えるだけの指示型インタラクションにとどまっており，stable diffusion[Rombach 22] をはじめとする画像生成インタラクションも指示型インタラクションの一種である．画像生成インタラクションは，ユーザが自身の意図を自然言語でモデルに伝えることで画像を生成するプロセスである．このプロセスでは，ユーザが脳内のイメージを言語化する際に，モデルが主観的な表現を理解できないという認識がギャップを引き起こす要因となっている．特に，生成画像が意図に合致しない場合，ユーザはプロンプトの作成や調整に多くの時間と労力を要することが課題である．従来の画像生成モデルは道具としての役割に特化しており，ユーザが明確で客観的な指示を与えることを前提としている．しかし，画像生成タスクはしばしば共創的な性質を持ち，ユーザが曖昧かつ感覚的な表現を用いることが多い．このような場面では，エージェントが単なる指示の受け手ではなく，ユーザの意図を推察し，補完する役割を果たすことが求められる．このようなエージェントの実現には,「主観表現理解の期待欠如を引き起こすエージェント性の欠如」という課題を解決する必要がある．これは，ユーザがエージェントに対してテキスト入力を行う際に，エージェントが感情を理解できないという見方を持つことによって，指示型インタラクションにみられるような客観的な指示文の作成を目指してしまうヒューマンファクタのことを指す．従来の LLM（大規模言語モデル）を利用したプロンプト補助では，表面的にはユーザの意図を反映しているように見えるが，実際にはモデルはユーザの内部状態や主

*連絡先：慶應義塾大学理工学研究科
〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail: matsuoka@ailab.ics.keio.ac.jp

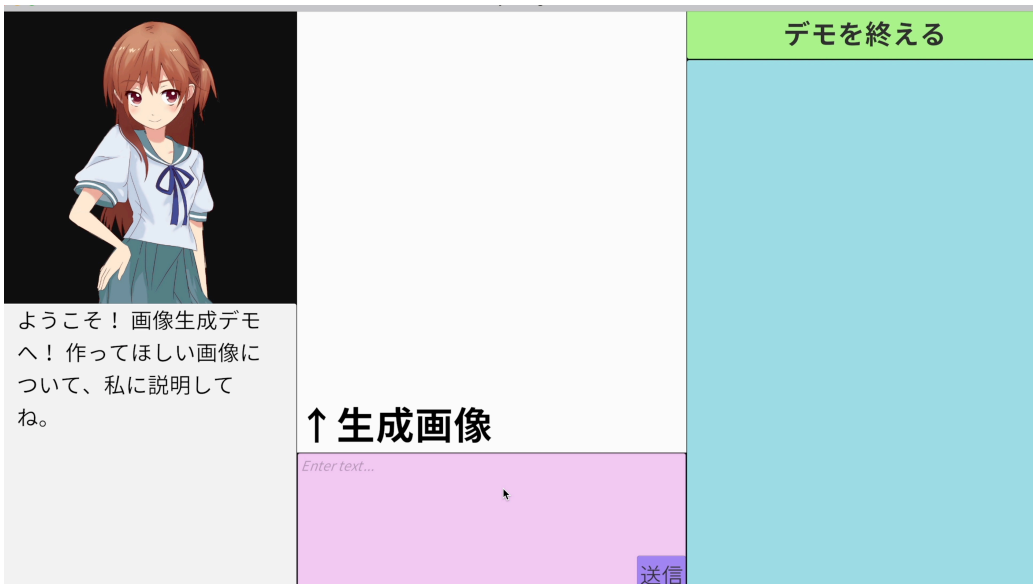

図 1: システム UI: テキスト入力条件

観的な意図を深く理解していない．このような状況では，ユーザがエージェントに対して「自分の感情を理解することはできない」という印象を抱きやすくなる．

Rostami らの研究 [Rostami 23] によれば，感情的に知能を持つチャットボットはユーザに共感されている感覚を提供する一方で，その信頼性や感情理解の限界が，ユーザにおける期待を阻害する要因となることが指摘されている．また，Svikhnushina らの研究 [Svikhnushina 20] によれば，人はエージェントに対して感情の理解を求めているものの，人が求めるものには至っていないとの結果が出ている．一方で，Schaaff らの研究 [Schaaff 23] では，GPT3.5 は自閉症患者よりも人の感情認識能力に長けていることがわかり，エージェントが感情認識に対して，ある程度の能力をすでに有していることがわかっている．

ヒューマンファクタの側面では，入力方法とエージェントの見た目の 2 つの要素がある．テキスト入力と音声入力の際については，Terblanche ら [Terblanche 23] によると，テキストの方が正確な伝達が可能であり，ユーザビリティに有意差はない．また，杉山ら [聡 98] によると，話し言葉は書き言葉に比べて少ないモーラ数での伝達が可能である．エージェントの見た目については，坂本らの研究 [坂本 07] によると，エージェントの見た目の問題は重要であり，エージェントが人に近い見た目をしているほど，人がエージェントを人らしいと判断する．これらのことから，音声対話エージェントとのインタラクションにより，主観表現理解の期待欠如の解決が見込める．

そこで，本研究では，エージェントへの入力形式としてテキスト入力と音声入力の違いが，ユーザの体験やエージェントのパフォーマンスに与える影響を調査することを目的とする．

2 関連研究

本研究の関連研究として，画像生成プロンプトの修正補助の研究及び，人同士のデザイン相互作用について述べる．

2.1 画像生成プロンプトの修正補助の研究

画像生成プロンプトの修正補助の研究としては，PromptCharm[Wang 24] や PromptPaint[Chung 23] のようにビジュアルイメージによる画像修正を混ぜ合わせる手法や，Promptify[Brade 23] のように複数画像を提示する方法，stable walk[Rost 23] にように，stable diffusion の画像空間を視覚的に表示することでプロンプティングを支援する手法といった研究があるが，これらの研究はユーザのプロンプト修正を支援することが目的である．また，Wen et al. のように，ハードプロンプトを学習させる [Wen 24] がある．これは，単語の意味解釈を固定する追加学習を画像生成モデルに行う手法である．これらの研究では，ユーザの試行錯誤の効率化を図ることを目的としており，ユーザとシステムのインタラクションは指示型のままである．

2.2 人同士のデザイン相互作用

Mace ら [Mace 02] は，創造における作業プロセスを 4 つの主要なフェーズに分けている．フェーズ 1 では

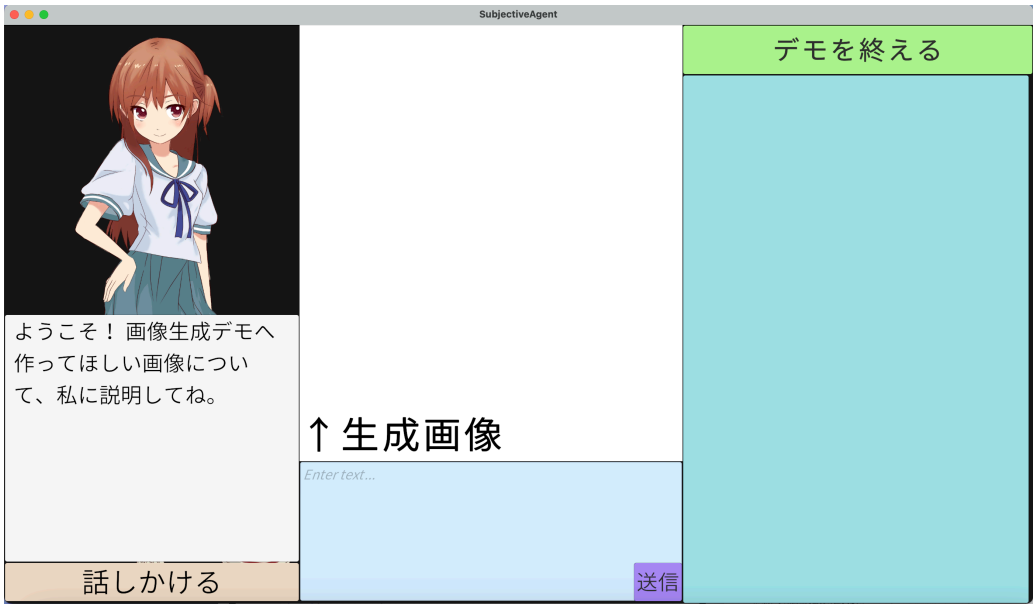

図 2: システム UI: 音声入力条件

作品の概念化が行われ，フェーズ 2 ではアイデアの展開に焦点が当てられる．フェーズ 3 では制作と改良が行われ，フェーズ 4 で作品の完成に至る．フェーズ 2 とフェーズ 3，さらにフェーズ 3 とフェーズ 4 の移行においては，視覚的な質や意図されたコンセプトとの整合性を評価し，アーティストが自身の作品を反復的に向上させるプロセスが含まれている．また，立原ら [立原 18] が，中学生 137 人に対して行った絵画の鑑賞調査では，主題として，語り的な「意図」を感受した生徒は 90 名に達し形容語的な「情感」はわずか 37 名に留まったことが示されており，この情感の感受性に個人差があることも示されている．

2.3 仮説

本研究では，画像生成時のエージェント性の有無とユーザ入力文の質に着目し，以下の仮説を立てた．

RQ1 エージェントに対して音声入力を行うと，テキスト入力時と比較して，主観的な文章が多くなる．

RQ2 音声入力の方がテキスト入力よりもワークロード負荷が小さい

3 設計

2.3 章で述べた仮説に基づき，以下の 2 条件をに応じたシステム設計を行なった

テキスト入力条件 ユーザはテキスト入力で画像説明を行う．

音声入力条件 ユーザは音声入力 (口頭) で画像説明を行う．

3.1 システム UI

本研究に用いたシステム UI は，図 1 および，図 1 のようになっており，図 1 はテキスト入力条件に対応し，図 1 は音声入力条件に対応したシステム UI である．システムは，左上にエージェントとして，響 [Live2D 24] の Live2D モデルを掲示し，すぐ下のボックス（セリフ掲示部）に，セリフが掲示される．真ん中の列は，上部に生成した画像を掲示するスペースがあり，下部はユーザの文章入力を行うスペースである．右側には，ユーザの入力履歴とエージェントが生成した画像がチャット形式で出力される．音声入力条件に適応した図 1 にのみ，UI 左下に橙色の「話しかける」ボタンがあり，ボタンによって音声入力の開始と終了をコントロールしている．また，図 1 でのみ，3.3 章に後述するシステムの発話を音声で読み上げる．読み上げには VOICEVOX[Hiroshiba 23] を用いた．

3.2 内部システム

内部システムは，松岡ら [松岡 24] のシステムをベースに，人の入力を文脈情報として保持し，文脈情報を人の入力として扱うようにした対話システムである．松岡らのシステムは，ユーザの入力を受けとると，画像を生成し，その画像に対して主観的な画像記述を付与する．付与した記述とユーザの入力を比較することで，

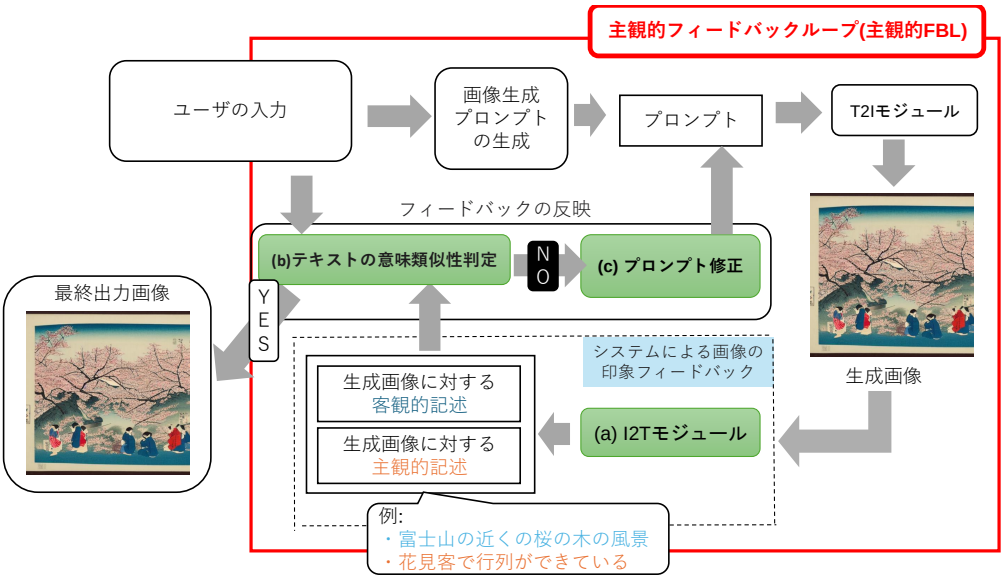

図 3: 内部システムのアーキテクチャ

主観情報を陽に扱うことができる．本研究では，ユーザの入力を単一のものではなく，会話履歴を要約して与えるようにすることにより，ユーザのコンテキスト情報を持ったインタラクションを可能にしている．

3.3 インタラクションの流れ

システムが起動すると，エージェントは「ようこそ！画像生成デモへ 作ってほしい画像について，私に説明してね.」と発話する．なお，この発話は 3.1 で述べた UI のセリフ掲示部に掲示される．ユーザが画像説明文を入力すると，エージェントが「了解！　今から作るからちょっと待っててね」と発話し，内部システムの処理が開始される．内部システムでは，図 3.2 の「テキストの意味類似判定部」において,「NO」と出力された場合，すなわち出力画像がユーザ入力を満たさないと判定された場合には，エージェントは,「うーん... ちょっと違うからもう一回作り直すね.」と発話し，生成した画像を提示した上でもう一度画像生成を行う．一方,「YES」と判定された場合には,「画像が生成されました！」とエージェントが発話し，最終出力画像が提示される．もし，最終出力画像に対して，ユーザが満足しなかった場合には，再度入力欄に画像説明を入力することで画像生成を再度行う．音声入力条件でのみ，エージェントの発話は，セリフ掲示部への掲示とともに音声出力が行われる．

4 実験

4.1 目的

本研究では，ユーザの入力文を分析し，2.3 章に提示した，以下に示す仮説 (RQ) について検証を行った．

RQ1 エージェントがいると，主観的な文章が発生する．

RQ2 主観的な文章を入れると，ユーザの意図と生成画像のギャップが少なくなる．

図 4: ケーススタディで提示した画像

4.2 実験手順

本実験では，20 代の男女 12 人を被験者とし，テキスト入力条件と音声入力条件を各 6 人ずつ実験し，被験

者間実験を行った．以下に条件の詳細について述べる．

テキスト入力条件 図 1 の UI を用いる．被験者は真ん中の青いテキスト入力欄からテキストを入力し，エージェントに指示を与える．

音声入力条件 図 1 の UI を用いる．被験者は，UI 左下のオレンジ色の「話しかける」ボタンを押して，指示を口頭で与える．

実験では，2 条件に共通して，図 4 示す画像を提示し，以下のように指示をした．

> この画像に似た画像をシステムに作ってもらいます．システムの指示に従って，画像を作成してもらってください．
> システムのピンク色の欄に，文章の入力を行い．終わったら送信ボタンを押してください．
> 生成された画像が思ったものと違った場合，違う点を指摘してください．
> 4〜5 回入力を行っても満足のいく画像が出なかった場合，その時点で終了してください．

ユーザはこの指示に従い画像タスクに取り組んだ．その後，アンケートに回答した．

4.2.1 アンケート項目

実験後に行うアンケートは，メンタルワークロード指標である NASA-TLX 日本語版 [芳賀 96] の 20 段階リッカートスケール評価および，自由記述のアンケートで構成される．自由記述形式のアンケートの質問項目を以下に示す．

1. システムに説明文を入力するときに苦労した点はどんなところでしたか？
2. システムの出力とあなたの意図にどれくらいギャップがありましたか？
3. システムにフラストレーションが溜まることはありましたか．具体的に教えてください．

4.2.2 主観表現の出現率

今回は絵画に対する主観表現の評価として，平均主観表現出現数 $P(\%)$ を計算した．P は，主観表現出現数を S，インタラクション回数を W として，式 1 のように示される．

$$P = \frac{S}{W} \times 100 \tag{1}$$

表 1: 主観表現の例示

単語	例文
画風に関するもの	「抽象画風の」「暗い雰囲気の」
量に関する記述	「たくさんの」「カラフルな」
比喩	「小学生が描いたみたいな」「ゴッホみたいな」

表 2: 音声入力条件とテキスト入力条件での平均主観出現数の差異*

条件	平均主観出現回数 (回)
音声入力条件	1.80
テキスト入力条件	1.11

なお，主観表現出現数 S は，表 1 のいずれかに当たる表現をカウントして算出した．本指標は，一人が 1 回のインタラクションにおける，平均の主観表現含有数を示すものである．

4.3 実験結果

実験では，ユーザは 4 回もしくは 5 回で入力を終えた．実験参加者 12 人中 5 人が 4 回で入力を終え，7 人が 5 回で入力を終えた．

4.3.1 RQ1 の検証

RQ1 の検証は，音声入力条件とテキスト入力条件における主観出現確率の差異についてついて有意である事を示す．表 2 は，音声入力条件とテキスト入力条件での平均主観出現確率の差異についての結果である．表 2 の結果を対応のない T 検定および，Cohen ' s d の効果量を測定したところ，$p = 0.02039261606(< 0.05)$ であり，Cohen ' s d の絶対値は 1.6 であった．よって，有意に音声入力の方が平均主観出現数が多く，効果量が多かったので，RQ1 は有意に音声入力が優位であると言える．

4.3.2 仮説 2 の検証

アンケートで取得した NASA-TLX の 5 項目を平均し，20 点満点の平均スコアを被験者ごとに計算し，2 条件間の平均スコアの差異を表 3 に示す．なお，20 に近いほど positive になるように，作業成績の項目以外を，スコアを 21 から引いて調整した．結果を対応のない T 検定および，Cohen ' s d の効果量を測定したところ，$p = 0.44(> 0.05)$ であり，Cohen ' s d の絶対値は 0.38 であった．よって，平均値ベースでは，テキスト

表 3: 音声入力条件とテキスト入力条件での NASA ー TLX スコアの差異

条件	平均スコア (20 点満点)
音声入力条件	12.58
テキスト入力条件	13.64

入力条件 ¿ 音声入力条件であったものの有意差は見られなかった.

4.3.3 アンケート結果からの課題

アンケート結果から，以下の課題が明らかになった.

- 語彙力不足による表現の難しさ: ユーザは自身のイメージを言語化する際に，適切な語彙を見つけることに苦労していた.
- システムとの意図のギャップ: 特に「タッチ」や「雰囲気」といった感覚的要素の伝達が難しく，結果として生成物が期待に合致しないことがあった.
- フラストレーションの原因: システムが以前の入力内容を忘れることや，意図した変更が反映されないことが，不満の原因となっていた.
- 音声認識結果のフィードバックで，自分の話した内容と異なる認識をした際にユーザはストレスを感じた.
- エージェントがやり直す際に発話することで，待たされている理由が明確に伝わるという意見があった．一方で，どこが違うのかまで知りたいという意見もあり，エージェントの判断基準の透明化が必要であるとわかる.

4.3.4 インタラクション回数と主観表現の関係

実験を行う中で，インタラクションの回数を重ねるとユーザが客観的な指示を出す場面が散見されたため，インタラクション回数と主観表現出現回数の関係について調べた．音声入力条件，テキスト入力条件についてそれぞれ反復測定 ANOVA による検定を行った．結果を表 4 に示す 音声入力条件では，$P = 0.75, F = 3.10x$ となり，P 値が 0.05 を大きく上回り，仮説は立証されなかった．テキスト入力条件でも，$P = 0.06, F = 3.10$ となり，P 値が 0.05 をわずかに上回ったが，極めて近いものにはなったので，有意傾向にあると言える.

表 4: 音声入力条件とテキスト入力条件での回数ごとの主観表現出現回数の差異

回数	音声入力条件	テキスト入力条件
1 回目	1.5	1.83
2 回目	2.17	1
3 回目	1.83	0.83
4 回目	1.83	0.6

5 まとめと展望

本研究では，人-エージェントによる画像生成インタラクションにおけるユーザの主観的な表現とインターフェースの差異の関係について調査した．音声認識とテキスト入力での主観言語の発露について，有意な差は出なかったものの，音声の場合の方が主観言語が多く見られた．アンケートでは，語彙力の不足をユーザが認識したケースが見られた．これは，テキストチャットにおいて見られた傾向であり，言語化のコストについてはさらに研究を進める必要がある．一方で，主観言語の多寡については，インタラクションの回数に関連がある可能性が示唆された．これは，人が伝わりにくいと感じた際に客観的な言葉を使い始めることにあると考えられる．インタラクション上の課題としては，エージェントが忘却することによるユーザのストレスが強いため，忘却しないことの重要性が示された．また，音声認識において，ユーザの発言が誤認識されたことが原因で，異なった画像が出された場合ユーザは余計な修正を強いられるようになった．その際，エージェントは「xx ではなく yy である.」といった言葉の置き換えの指示を実行することが不可能であった．音声インタラクションでは，音声認識のミスを完全に防ぐことはできないので，言葉の置き換えに対応するエージェントの研究が必要である.

謝辞

本研究は，JST，CREST，JPMJCR19A1 の支援を受けたものである.

参考文献

[Brade 23] Brade, S., Wang, B., Sousa, M., Oore, S., and Grossman, T.: Promptify: Text-to-image generation through interactive prompt exploration with large language models, in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–14 (2023)

[Chung 23] Chung, J. J. Y. and Adar, E.: Promptpaint: Steering text-to-image generation through paint medium-like interactions, in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17 (2023)

[Hiroshiba 23] Hiroshiba, K.: VOICEVOX - 無料で使える中品質なテキスト読み上げソフトウェア (2023), Accessed: 2025-01-24

[Live2D 24] Live2D, : Live2D サンプルデータ集（無料配布） (2024)

[Mace 02] Mace, M.-A. and Ward, T.: Modeling the creative process: A grounded theory analysis of creativity in the domain of art making, *Creativity Research Journal*, Vol. 14, No. 2, pp. 179–192 (2002)

[Rombach 22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022)

[Rost 23] Rost, M. and Andreasson, S.: Stable Walk: An interactive environment for exploring Stable Diffusion outputs, in *Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023)*, Vol. 3359 of *CEUR Workshop Proceedings*, pp. 89–97, CEUR-WS.org (2023)

[Rostami 23] Rostami, M. and Navabinejad, S.: Artificial Empathy: User Experiences with Emotionally Intelligent Chatbots, *AI and Tech in Behavioral and Social Sciences*, Vol. 1, No. 3, pp. 19–27 (2023)

[Schaaff 23] Schaaff, K., Reinig, C., and Schlippe, T.: Exploring ChatGPT's Empathic Abilities (2023)

[Svikhnushina 20] Svikhnushina, E. and Pu, P.: Social and Emotional Etiquette of Chatbots: A Qualitative Approach to Understanding User Needs and Expectations (2020)

[Terblanche 23] Terblanche, N. H. D., Wallis, G. P., and Kidd, M.: Talk or Text? The Role of Communication Modalities in the Adoption of a Non-directive, Goal-Attainment Coaching Chatbot, *Interacting with Computers*, Vol. 35, No. 4, pp. 511–518 (2023)

[Wang 24] Wang, Z., Huang, Y., Song, D., Ma, L., and Zhang, T.: PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement, in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–21 (2024)

[Wen 24] Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., and Goldstein, T.: Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery, *Advances in Neural Information Processing Systems*, Vol. 36, (2024)

[坂本 07] 坂本大介, 神田崇行, 小野哲雄, 石黒浩, 萩田紀博：遠隔存在感メディアとしてのアンドロイド・ロボットの可能性, 情報処理学会論文誌, Vol. 48, No. 12, pp. 3729–3738 (2007)

[松岡 24] 松岡竜輝, 熊野史朗, 今井倫太, 成松宏美：意味的類似性にもとづく主観的印象テキストからの画像生成, 人工知能学会 言語・音声理解と対話処理研究会（SLUD）第 102 回研究会 (第 15 回対話システムシンポジウム), pp. 223–228 (2024)

[聡 98] 聡 杉山, 浩二 堂坂, 豪 川端：話しことば対話によるテキスト内容の伝達, 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 21, pp. 83–90 (1998)

[芳賀 96] 芳賀 繁, 水上 直樹：日本語版 NASA-TLX によるメンタルワークロード測定, 人間工学, Vol. 32, No. 2, pp. 71–79 (1996)

[立原 18] 立原 慶一：鑑賞能力と主題感受の関わりを求めて, 美術教育学：美術科教育学会誌, Vol. 39, pp. 209–222 (2018)