フォーカシング対話の収集と LLM を用いた EXP スケールの自動評定

江 舒婷 [1] 郭 傲 [1] 青木 剛 [2] 中西 美和 [3] 東中 竜一郎 [1]
[1] 名古屋大学 [2] 龍谷大学 [3] 南山大学
jiang.shuting.g7@s.mail.nagoya-u.ac.jp, guo.ao.i6@f.mail.nagoya-u.ac.jp,
aokit@psy.ryukoku.ac.jp, miwanaka@nanzan-u.ac.jp, higashinaka@i.nagoya-u.ac.jp

概要

近年，精神疾患の治療だけでなく，日常的なメンタルヘルスケアの重要性が高まっている．こうしたケアを広く提供するために，チャットボットを用いた支援が期待されているが，有効な支援を行えるものは未だ確立されていない．その要因として，対話データが不足している点や，日常的なメンタルヘルスケアの対話の質を自動的に推定する仕組みが未確立である点が挙げられる．そこで本研究では，上記の二つの課題に対処するため，フォーカシングに基づく心理カウンセリング対話データを収集した．さらに，収集したデータを用いて，LLM による，フォーカシングにおける対話の質の指標である EXP スケール自動評定の実現可能性を検証した．実験の結果，LLM は EXP スケールを一定の精度で推定できることを確認した．

1 はじめに

近年，日常的なメンタルヘルスケアの重要性が高まっており [1]，その重要性は精神疾患の治療に劣らない．このようなケアを広く提供するため，チャットボットなどの対話システムによる自動化支援が期待されている [2]．しかし，既存のシステムの多くは表面的な対話にとどまり，ユーザの深い内省を効果的に促すシステムはまだ確立されていない．

このようなシステムが実現されていない主な理由は，開発に必要な対話データの不足である．既存のデータセットの多くはうつ病診断や自殺予防などの「治療」場面を対象としており，日常的なストレス管理や心理的ウェルビーイングの向上を目的としたデータセットは極めて少ない [3]．

日常的なケアに有効な心理療法として「フォーカシング」 [4, 5] が注目されている．フォーカシングは，身体感覚への注意を促し，内的な気づきを導く技法である．自己理解を深め，心理的課題の解決を図ることができる．このプロセスを支援する対話システムが実現すれば，ユーザは日常のあらゆる場面でセルフケアが可能となり，主観的ウェルビーイングの向上に大きく寄与すると期待される．しかしながら，そのような対話システムの構築には，少なくとも二つの課題がある．一つ目は，学習や評価に用いうるフォーカシング対話データが十分に収集・整備されていないことである．二つ目は，仮にそのようなシステムを構築できたとしても，その対話の質を自動的に推定する仕組みが未確立であることである．

本研究では，上記の二つの課題に対処するため，フォーカシングに基づく心理カウンセリング対話データセットの収集を行う．また，収集したデータを使用して，LLM による，フォーカシングにおける対話の質の指標である EXP スケールの自動評定実験を実施し，LLM が EXP スケールをどの程度推定できるかを検証する．

2 フォーカシングと EXP スケール

本節では，フォーカシングの概念と，本研究で導入するイメージワーク，および，EXP スケールの概要について述べる．

2.1 フォーカシング

フォーカシングは Gendlin [6] によって創始された療法であり，その核心は「フェルトセンス」と呼ばれる曖昧な身体感覚に注目し，それを言葉やイメージとして「象徴化」することで，心身の解放をもたらす「フェルトシフト」を誘発することにある [7]．これは生理的なリラックスをもたらすだけでなく，心理的問題に対する新たな視点も提供する．

2.2 イメージワーク

フォーカシングにおいて，イメージワークは単なる派生的技法ではなく，体験過程理論の核心に根ざした

アプローチである．フォーカシング創始者の Gendlin は，フェルトセンスを同定する「ハンドル」として，言語と同様にイメージが極めて有効であることを指摘している [7]．特に身体感覚の言語化に不慣れな人にとって，内面を何もない状態から言語化することは認知的負荷が高い．イメージワークを用いることで，フォーカシング療法の効果を大きく高めることができる．

本研究では，イメージワークとして，「心の天気」と「アニマルクロッシング」を用いる [8]．「心の天気」は心境を天気（晴れ，雨など）に投影し，「アニマルクロッシング」は動物のイメージを借りて内的状態を具体化する．このようなイメージワークは，曖昧な感覚から明確な象徴への移行を加速させ，フェルトシフトを効果的に促進し，クライエント（来談者；カウンセリングの対象者）が認知的反芻から抜け出し，自己受容を高めることを支援する．

2.3 体験過程尺度（EXP スケール）

体験過程尺度（Experiencing Scale；以下 EXP スケール）は，クライエントが自身の内的体験に接触・探索し，それを表現する深度を評価する指標である．フォーカシングは，浅い認知レベルから深い身体感覚（フェルトセンス）への移行を促進する技法であるが，EXP 尺度は，まさにこの「体験の深化」を測定するために開発された尺度である．

本研究では，池見ら [9] によって開発された標準的な日本語版 EXP スケールを採用する．表 1 に各段階の評価基準を示す．EXP スケールを用いることで，フォーカシング対話における内的探索の深化を定量的に捉えることができる．また，本尺度を用いることで自動評価も可能となる．

近年の生成 AI 技術の進展は，LLM を用いた EXP の自動評定という新たな解決策を提示している．先行研究として，福盛ら [10] は，一般的なロジャーズの面接記録を対象に GPT-5 による自動評定を検証し，専門家と中程度の一致（重み付きカッパ係数 0.42）を確認した．しかし，その一方で，AI の評定傾向やプロセスの不透明性といった課題も報告している．また，Yamagata ら [11] は GPT-4o および o1 モデルを用い，3 段階評定において高い一致度（カッパ係数 0.79〜0.94）を達成し，スクリーニングツールとしての有用性を示唆した．しかし，イメージワークを含む対話文脈において，LLM がどの程度正確に体験の深さを判定できるかは未だ十分に検証されていない．

3 データ収集とアノテーション

本節では，本研究のために独自に構築したフォーカシング対話データセットの収集および整備のプロセスについて述べる．以下，収録環境，実験手順，および専門家によるアノテーションの詳細を順に説明する．

3.1 機材・実験参加者

本実験の収録環境を図 1 に示す．言語情報に加え，身体動作や表情などの非言語情報を多角的に記録するため，以下の機材で音声と映像を同期収録した．

音声収録　音声を高品質に個別収集するため，五つのマイクアレイを搭載している首掛け式ウェアラブルマイク（Thinklet[1)]）を計 2 台使用した．各参加者に装着し，各話者の音声を個別に収録できる．

映像収録　参加者の表情と身体動作を記録するため，広角アクションカメラ（GoPro）を計 3 台使用した．

- **GoPro ①・②**：机上に配置し，カウンセラーとクライエントの表情と上半身の動作を正面から撮影する．
- **GoPro ③**：三脚に取り付け，両参加者の側面を第三者視点から撮影する．

データ統合・制御　全機器（計 5 台）を中央 PC に接続し統一管理した．また，カウンセリング対話の自然性を最大限保証するため，専用の制御プログラムを開発した．これにより，カウンセラーは 1 つのインターフェースのみで全機器の録画開始と終了を制御でき，研究者はカウンセリング開始前に退室できる．この設計により，第三者の在室や録画機器がクライエントに与える心理的影響を効果的に軽減し，録画を過度に意識することなく，リラックスした状態で自然に対話できる．

実験参加者については，南山大学と名古屋大学からクライエント（男性 2 名，女性 6 名，平均年齢 21.5 歳）を募集した．カウンセラーは 2 名の専門家（本研究の共著者）が担当した．両名とも公認心理師および臨床心理士の資格を有し，フォーカシング指向心理療法における豊富な臨床経験を持つ．

参加者は 2 名 1 組で対話セッションを行った．カウンセラーは「こころの天気」および「アニマルクロッシング」のイメージワークを通じて，クライエントの内的状態の表現を促した．実験開始前に，研究者は全

1) https://mimi.fairydevices.jp/technology/device/thinklet/

表 1　7 段階 EXP スケールの概要 [12]

段階	評定基準
段階 1	話し手と関連のない外的な出来事について語る．
段階 2	話の内容は話し手と関連があるが，話し手の感情は表明されない．知的・行動的な自己描写にとどまる．
段階 3	外的出来事に対する感情は表明されるが，そこから自分自身については述べない．
段階 4	出来事に対する体験や感情を話題とし，自分の体験へ注意を向けて膨らませたり深めたりする．
段階 5	自分の抱える問題に対して，問題や仮説を提出する．探索的・思考的・ためらいがちな話し方が特徴である．
段階 6	新しい感情や体験に新たに気づく．新しい自己体験や感情の変化について話す．
段階 7	感情や内的過程についての気づきが拡がっていく．

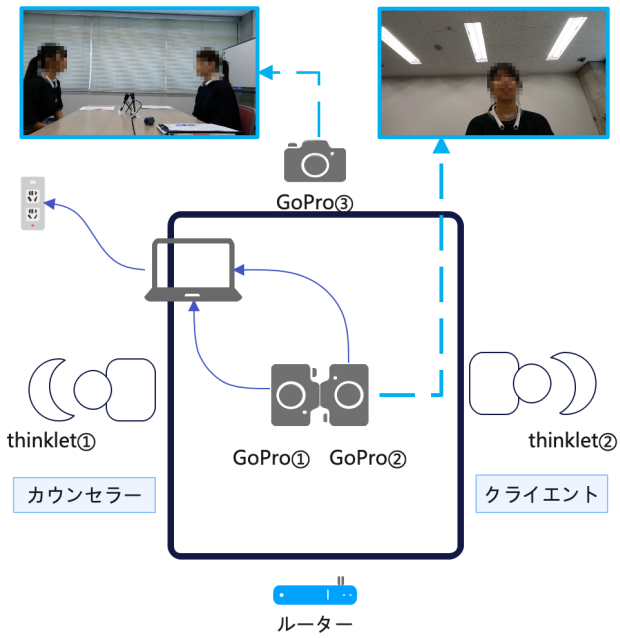

図 1　実験環境の構成図

参加者に研究目的，手順，音声・映像データの処理方法，プライバシー保護措置について書面および口頭で説明し，書面によるインフォームド・コンセントを取得した．本研究は所属組織における倫理委員会の承認を得ている．

3.2　実験手順

実験環境の調整とインフォームド・コンセントを経た後，参加者に Thinklet 等の機材を装着させ，収録体制を整えた． 実験セッションでは，約 20 分間のフォーカシング対話を実施し，終了後にはアンケート，および，振り返りインタビューを行った．

3.3　データ整備

データ収集後，まずタイムスタンプに基づいて Thinklet で録音した音声を同期し，2 チャンネルの音声ファイルを生成した．次に，専門スタッフが厳密な逐語録を作成した．文字起こしでは，相槌やフィラーも忠実に記録し，発話区間の開始時刻と終了時刻を正確に標定した．構築したコーパスの統計量を表 2 に示す．

表 2　データセットの統計

項目	内容 / 統計
参加者構成	カウンセラー 2 名，クライエント 8 名
総セッション数	8 セッション
総収録時間	196.93 分（平均 24.6 分/セッション）
モダリティ	2ch 音声，テキスト（書き起こし）
総発話数	3,596 発話 (平均 499．5 発話/セッション)
総単語数 *	30,050 語
平均発話長	8.36 語/発話

* 形態素解析器 MeCab [13] を用いて算出

3.4　EXP のアノテーション

アノテーションソフトの ELAN [14] を用い，発話単位を最小分析単位として，EXP スケールのアノテーションを実施した．分割された長文についてはひとまとまりとしてアノテーションを行った．EXP スケールの評定は，2 名の心理学専門家（本研究の共著者）が実施した．うち 1 名は臨床心理士および公認心理師の資格を持ち，国際フォーカシング研究所（TIFI）認定のフォーカシング専門家資格を取得している．評定の信頼性を確保するため，2 名の評定者は独立して評定を行った後，不一致となったケースについて議論し，最終的な EXP スコアを確定した．

本データにおける EXP スコアの分布について述べると，段階 3（47.57%）と段階 4（36.15%）の合計が 8 割以上（83.7%）を占めており，データが主に「内的感情の表現（段階 3)」から「体験へのフォーカシング（段階 4)」に集中していることがわかる．この傾向は，クライエントが外的出来事の記述レベルに留まらず，自身の感情や感じられる意味に触れる探索的な姿勢を

表 3　先行研究と本手法の自動評価結果．SD は標準偏差．↓は値が小さいほど，↑は値が大きいほど性能が良いことを示す．

手法	自動評価値		人手評価値		評価指標					
	平均	SD	平均	SD	MAE↓	適合率↑	再現率↑	単純一致↑	カッパ値↑	順位相関↑
Yamagata らの手法	2.53	1.22	3.40	0.66	1.23	0.23	0.19	0.26	0.19	0.29
本手法	**3.65**	**1.01**	3.40	0.66	**0.72**	**0.36**	**0.39**	**0.42**	**0.34**	**0.40**

保持していることを示している．

4　EXP 自動評定実験

前節で構築したデータセットを用いて，LLM による EXP スケール自動評定の有効性を検証した．

4.1　プロンプト

本実験では OpenAI 社の GPT-5 モデルを採用した．同モデルは強力な推論能力と文脈理解能力を持つとされる．EXP 評定の精度向上のため，以下の 4 要素を含むプロンプトを設計した：

- **専門家役割設定**：システムプロンプトで LLM に「ロジャーズの人間中心療法と EXP スケールに精通した臨床心理学専門家」の役割を付与し，臨床心理学的視点に基づく解釈を促す．
- **対話手順と対話履歴**：EXP 評定は文脈依存であるため，対象発話に加えて，イメージワークの対話手順と全体の対話ログを入力する．
- **Few-shot 学習**：各段階の定義に加え，我々が事前に行った予備実験から得られた対話事例を複数入力し，抽象的定義では捉えにくい微妙な基準をモデルに与える [15]．
- **Chain-of-Thought**：判断根拠を先に生成してから評定スコアを出力することで，推論の精度と透明性を高める [16]．

4.2　比較手法

本研究では，Yamagata らの論文中で最も高い性能を示した EXP 自動評定プロンプトを比較手法として採用し，本手法の有効性を検証する．Yamagata ら [11] の研究は，公開されている心理学の書籍やマニュアルに記載された典型的な事例を対象としてプロンプトを作成しており，実験の結果，EXP スケールの「要約版定義」と「Few-shot」を組み合わせたプロンプトが最も高い性能を示したと報告されている．一方で，本研究の「イメージワーク」を含む対話に対しては未検証である．

4.3　評価指標

モデルの評定性能を多角的に評価するため，単純一致率，重み付きカッパ係数，スピアマン順位相関係数，平均絶対誤差（MAE）の 4 つの定量指標を採用する．正解ラベルには 3.4 節で述べた専門家合議スコアを使用する．

4.4　実験結果

表 3 に結果を示す．Yamagata らの手法では，重み付きカッパ係数が 0.19，順位相関が 0.29，MAE が 1.23 と比較的大きな誤差が生じている．これは，Yamagata らの手法がフォーカシング療法特有の「体感的意味」への言及や，イメージワークを通じた微妙な体験の深化を十分に捉えられなかったためと考えられる．

一方，本手法では，重み付きカッパ係数が 0.34，単純一致率が 0.42 となるなど，すべての指標で Yamagata らの手法を上回った．特に MAE は 0.72 まで低下し，予測のずれが大幅に減少した．これは，専門家役割の付与と詳細な文脈情報の提供により，LLM が対話の流れや発話意図をより正確に解釈できるようになったことを示している．

5　まとめ

LLM による EXP スケールの自動評定の実現可能性を探るため，本研究ではフォーカシングとイメージワークに基づく心理相談対話データセットを構築し，プロンプトを設計して評定実験を実施した．先行研究の手法と比較して，本手法は高い性能を示した．また，絶対誤差が比較的小さいことから，この結果は，LLM に基づく EXP スケールの自動評定が一定の一致性レベルに到達したことを示していると言える．今後の研究では，データセットの規模を拡大し，さらに音声・映像などのマルチモーダル情報を活用することで，評定精度のさらなる向上を図りたい．また，フォーカシングを行う音声対話システムの開発を進めたい．

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けた．

参考文献

[1] Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. No health without mental health. **The Lancet**, Vol. 370, No. 9590, pp. 859–877, 2007.

[2] Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. **Frontiers in Digital Health**, Vol. 5, p. 1133987, 2023.

[3] Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. KokoroChat: A Japanese Psychological Counseling Dialogue Dataset Collected via Role-Playing by Trained Counselors. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 12424–12443, 2025.

[4] Eugene T Gendlin. Focusing. **Psychotherapy: Theory, Research & Practice**, Vol. 6, No. 1, p. 4, 1969.

[5] Eugene T Gendlin. **Focusing-oriented psychotherapy: A manual of the experiential method**. Guilford Press, 2012.

[6] Marjorie H Klein, Philippa Mathieu-Coughlan, Eugene T Gendlin, and Donald J Kiesler. **The experiencing scale: A research and training manual**. Wisconsin Psychiatric Institute, 1969.

[7] Eugene T Gendlin. **Focusing**. Bantam Books, 1981. Revised edition.

[8] 池見陽. 心のメッセージを聴く. 講談社, 1995.

[9] 池見陽. 体験過程とその評定：EXP スケール評定マニュアル作成の試み. 人間性心理学研究, Vol. 4, No. 3, pp. 50–64, 1986.

[10] 福盛英明, 永野浩二, 青木剛, 森川友子, 竹田悦子, 平野智子. EXP スケールの評定者の評定経験の特徴と自動評定との比較に関する探索的検討. 日本人間性心理学会第 44 回大会発表論文集, 2025.

[11] Midoriko Yamagata, Daisuke Yamagishi, and Akira Ikemi. Rating the Experiencing Scale with generative AI: examining the accuracy, reliability, and feasibility of ChatGPT. **Person-Centered & Experiential Psychotherapies**, pp. 1–16, 2025.

[12] 久保田進也・池見陽. 体験過程の評定と単発面接における諸変数の研究. 人間性心理学研究, Vol. 9, pp. 53–66, 1991.

[13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[14] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: A professional framework for multimodality research. In **Proceedings of the 5th International Conference on Language Resources and Evaluation**, pp. 1556–1559, 2006.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In **Proceedings of the 36th International Conference on Neural Information Processing Systems**, pp. 24824–24837, 2022.