カウンセリングドメインに特化して fine-tuning を行った LLM に対する Active Listening Skill の評価

三浦拓人 [1]　Natthawut Kertkeidkachorn[1] 小島治幸 [2]　白井清昭 [1]
[1] 北陸先端科学技術大学院大学 先端科学技術研究科　[2] 金沢大学 人間社会研究域
[1]{s2460005,natt,kshirai}@jaist.ac.jp
[2]hkojima@staff.kanazawa-u.ac.jp

概要

近年，カウンセリングに特化した Large Language Model(LLM) の開発が盛んに行われ，様々な観点からの評価が必要となっている．我々は，相手の話を理解していることを示す能力である Active Listening Skill(ALS) に着目する．カウンセラーの ALS は患者の自己開示を促すなど，カウンセリングにとって重要である．本研究では，カウンセリング対話で fine-tuning した LLM を基にカウンセリングチャットボットを作成し，ALS に関連するいくつかの評価項目によって当該ボットを評価した．評価の結果，一般的な LLM と比べて，我々が作成したモデルは言い換えや要約による ALS によって，自己開示をユーザに促すことができることを示した．

1　はじめに

カウンセリングとは，専門知識や技術を持つカウンセラーが相談者の悩みや困りごとの解決や支援を目的とした対話のことである．最近の調査では，カウンセリングに対する関心や需要が高まる一方で，心理的な抵抗，費用や時間などの制限などの障壁から，多くの人々がカウンセリングを受けていない現状が明らかとなった [14]．これらの障壁が低いカウンセリングチャットボットは，カウンセリングを受けることに躊躇がある人やカウンセリングを受けることができない人の助けに繋がることから開発が盛んに進められている．これに伴い，カウンセリング対話に必要な知識や能力を Large Language Model(LLM) に獲得させることを目指した研究が盛んに行われている [13, 7, 18]．カウンセリングは非常に複雑で難しいシナリオであるため，カウンセリングドメインに特化した LLM の開発を行う際には様々な観点からの精細な評価や効果測定が重要となる．

本研究では，カウンセリングの中でも重要なスキルである Active Listening Skill(ALS) があまり評価されていないことに着目する．ALS は，相手の発話の言い換えや要約，共感的な応答などによって，相手の話を理解していることを示す能力である [3]．適切な ALS は，効果的なカウンセリングを行うにあたって必要不可欠な要素である [15, 8, 9]．患者自身の自己理解やカウンセラーに対する患者の自己開示を促すためである．我々は，カウンセリング対話に関するコーパスを利用して LLM を fine-tuning し，そのモデルをベースとしたカウンセリングチャットボットを開発した．そして，ALS に関するいくつかの観点から，開発したチャットボットを評価した．

本論文の構成は以下の通りである．2 節では，ALS に関する対話システムの開発を目指した関連研究と本研究の位置付けについて述べる．3 節では，我々が開発したカウンセリングチャットボットの詳細について説明する．4 節では，このカウンセリングチャットボットを評価するために行った実験について述べる．最後に，5 節で本論文をまとめる．

2　関連研究

Chaszczewicz らは，LLM を活用して，初級カウンセラーの訓練のために，第 3 者の観点からカウンセリング応答に関するフィードバックを与えるアプローチを提案した [2]．上級心理療法専門家のグループと共同設計して，フィードバックのマルチレベル分類を開発した．また，400 件のカウンセリング会話に対して，包括的なフィードバックアノテーションを行い，そのデータセットを公開した．このデータセットには，基本的なアクティブリスニングの原則に沿っているかをフィードバックするためのアノテーションも含まれている．しかし，ALS の各

図 1　チャットボットのインターフェース

要素を細かく分類し，カウンセリング対話に対してそれらの細かい要素を評価することは行われていない．我々は，言い換え/要約/共感という観点からALS に関して細かく評価を行う．

Demasi らは，自殺防止ホットラインのカウンセラーを訓練するために，LLM による仮想の患者によってシミュレーション訓練を行うフレームワークを提案した [4]．シミュレーション訓練によって得られた会話に対して，ALS，de-escalation 戦略，社会的規範などの 25 の評価を行い，アノテーションした．この研究は，実際の人間の患者を対象としているわけではない．また，実際にカウンセリングを受けた人間によってカウンセラーの ALS の評価が行われていない．我々は，実際の人間の被験者を患者役としてカウンセリング対話を行う実験を実施し，カウンセリングチャットボットの ALS を被験者によって評価した．

3　カウンセリングチャットボット

3.1　カウンセリングドメインに特化した LLM の fine-tuning

本研究では，一般的な対話データセットによって訓練された LLM をベースとし，カウンセリングに関する対話データセットによる fine-tuning を行うことで，カウンセリングドメインに特化した LLM を開発した．我々は，ベースの LLM として Llama[16] を採用した．Llama-3-8B[1] に対して，日本語の会話データセット[2] で追加訓練が行われたモデル[3] を使用した．

我々は，上記の LLM に対して，2 つのデータセットを用いて fine-tuning を行った．1 つは，人間とチャットボットが人生の問題について話し合う英語の会話データセット[4] を日本語に翻訳することで，新たにデータセットを作成したものである．これは，99,469 会話が収録され，約 10 ターン (発話と応答) で 1 つの会話が構成される．もう 1 つは，カウンセラーと患者の実際のカウンセリング対話を収集した PDF ファイル [19] からデータを抽出したものである．これは，3 つの会話が収録され，それぞれ 92，62，17 のターン (発話と応答) で構成される．

これらのデータセットを利用して，表 1 に記載された訓練設定により，fine-tuning を実施した．計算資源の都合上，LoRA[10] を使用して fine-tuning を行った．我々が fine-tuning を行ったモデルを

1）https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
2）https://huggingface.co/datasets/fujiki/japanese_hh-rlhf-49k
3）https://huggingface.co/haqishen/Llama-3-8B-Japanese-Instruct
4）https://huggingface.co/datasets/jerryjalapeno/nart-100k-synthetic

表 1　fine-tuning の訓練設定

epoch	2
batch size	1
learning rate	2e-4
weight decay	0.001
warmup ratio	0.03
gradient accumulation steps	16
max sequence length	512

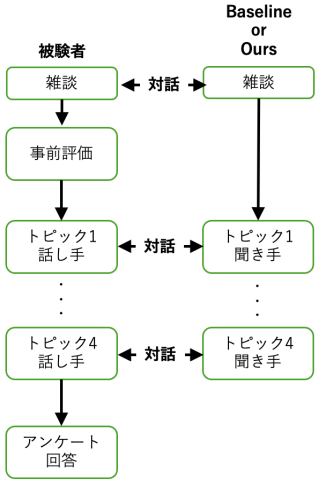

図 2　評価実験の流れ

huggingface 上で公開している[5]．

3.2　カウンセリングチャットボットの開発

我々は，前述したカウンセリング対話に特化して fine-tuning を行った LLM を基に，カウンセリングチャットボットを作成した．具体的には，機械学習モデルのデモを行う Web アプリケーションを GUI 上で作ることができる Python のライブラリ Gradio[6] を利用した．作成したチャットボットは，ユーザの発話を起点として，ユーザとシステムが 1 回ずつ交互に発話を行うことで，テキストベースの対話を実施する．図 1 で示すように，モデルに与える prompt や応答生成時のハイパーパラメータを GUI 上で操作できる仕様である．

4　Active Listening Skill の評価実験

4.1　実験設定

本研究では，チャットボットによる ALS の効果測定を行った研究 [17] を参考に，図 2 に示すような評価実験を設計した．具体的には，カウンセリングに関わる 4 つの質問トピックについて，それぞれ 10 回のやり取り (発話と応答) が行われるインタビュー対話を実施した．インタビュー対話の設定については，被験者を話し手，チャットボットが聞き手，すなわち被験者が患者，チャットボットがカウンセラーとなることを想定した．カウンセリング対話に特化した LLM(以下 LLM(fine-tune) と記す) に対する比較対象として，カウンセリングドメインの fine-tuning を行っていないベースの LLM(LLM(pre-train) と記す) を採用した．被験者には，2 つのモデルのうち，どちらか 1 つとのみ対話を行ってもらった．実験には，26 名の日本人大学生 (19 歳〜28 歳：男性 10 名，女性 16 名) が被験者として参加した．被験者は 2 つのモデル群のいずれかに無作為に振り分けられた．全てのトピックについて対話を行った後，ALS に関する評価を行うためのアンケートに回答するよう被験者に依頼した．また，チャットボットとの対話や対話システムの利用に慣れてもらうことを目的として，被験者にはインタビュー対話を始める前にチャットボットと雑談を実施してもらった．さらに，チャットボットとの対話に対する被験者のモチベーションを事前に調査した．

Xiao らが実施した実験 [17] を参考に，カウンセリングに関わる 4 つの質問トピックは，以下とした．

1. 2–3 文であなた自身について教えてください。
2. 暇な時には何をして楽しんでいますか？
3. あなたの一番良いところは何ですか？
4. 現在直面している最大の課題は何ですか？

モデルに対して「あなたは親切で楽しい精神療法助手です。(トピック文) という質問について、インタビュー会話を行ってください。」という prompt を与えて，被験者が該当の質問トピックに回答するところから対話をスタートするよう設定した．

4.2　評価項目

本研究では，カウンセリングチャットボットに対する ALS の評価項目として，次の 4 つの指標を設定した．

- チャットボットはあなたの話をどの程度理解していると感じましたか？ (総合評価)
- チャットボットはあなたの話をどの程度適切に言い換えられていましたか？ (言い換え評価)

5）　https://huggingface.co/bbookarisara/arisara_llama3_8b_for_mental_therapy
6）　https://www.gradio.app/

表 2　実験結果：被験者の平均スコアと標準偏差 (右下の数値)

モデル	事前評価	ALS 評価				印象評価		
	モチベーション	総合	言い換え	要約	共感	自己開示	信頼感	親密感
pre-train	$3.31_{0.95}$	$2.77_{1.24}$	$3.46_{1.56}$	$3.08_{1.71}$	$4.08_{2.02}$	$4.08_{1.50}$	$2.15_{0.99}$	$2.69_{1.44}$
fine-tune	$3.92_{0.75}$	$2.85_{0.99}$	$3.92_{1.04}$	$3.31_{1.11}$	$3.15_{1.52}$	$5.08_{1.44}$	$2.46_{1.27}$	$2.62_{1.19}$

- チャットボットはあなたの話をどの程度適切に要約していましたか？ (要約評価)
- チャットボットはあなたの話に対してどの程度共感してくれていると感じましたか？ (共感評価)

言い換え，要約，共感は ALS を構成する代表的な要素であり [3]，総合評価とは別にそれぞれを個別に評価する．また，チャットボットが有する ALS の効果測定を行うために，次の 3 つの指標も設定した．

- 対話システムに対してどの程度自己開示できましたか？ (自己開示)
- 対話システムをどの程度信頼できましたか？ (信頼感)
- 対話システムに対してどの程度親しみを覚えましたか？ (親密感)

カウンセリングにおける ALS の大きな意義として，ユーザからの自己開示を促し，効果的な治療へと繋げることが挙げられる [11, 6, 12]．また，カウンセリングにおいて ALS の効果として，ユーザと信頼感や親密感を築くことが挙げられる [1, 5]．これらの評価項目のそれぞれに対して，7 段階のリッカートスケールで回答するよう被験者に依頼した．被験者のモチベーションを調査する事前評価では，「チャットボットと雑談をしてみて、今から始まるインタビュー対話にどの程度興味が湧きましたか？」という項目を設け，前述の評価項目と同様，7 段階のリッカートスケールで回答するよう被験者に依頼した．

4.3　実験結果

実験の結果を表 2 に示す．2 群間で事前のモチベーションに統計的に有意な差はなかった．また，ALS 評価と印象評価の全ての項目で統計的に有意な差は見られなかった．ALS 評価では，共感以外の項目で fine-tuning 後のモデルが優れた評価を得た．共感において fine-tune が pre-train よりも低い評価を得た原因として，fine-tuning に利用した対話コーパスに共感的なやり取りが少ないからであると予想され，今後の詳細な調査が必要であると考える．

印象評価では，自己開示と信頼感の項目で fine-tuning 後のモデルが優れた評価を得た．一方で，親密感は，優れた評価を得られなかった．カウンセリング対話による fine-tuning を行ったことで，カウンセラーという役割を踏まえて丁寧な対話を行う傾向が強くなり，フランクさやカジュアルさが失われたことが要因であると考察する．

全体的な結果から，通常の LLM と比べて，カウンセリングドメインによる fine-tuning 後のモデルは言い換えや要約によって ALS を示すことにより，自己開示をユーザに促すことができる能力を有していると言える．本実験は，カウンセリングコーパスによる fine-tuning が，良好なカウンセリング対話システムの開発に繋がる可能性を示した．

5　おわりに

本研究では，カウンセリングの中でも重要なスキルである Active Listening Skill(ALS) の観点からこのチャットボットの有効性を評価した．実験の結果，一般的な LLM よりもカウンセリングドメインに特化した LLM の方が優れた ALS を有していることを示した．また，患者との信頼感を形成し，自己開示を促す効果があることを示した．

今後は，今回の実験で収集した対話データに対して ALS に関する評価項目をアノテーションすることで，ALS 評価ラベル付きカウンセリング対話コーパスを公開することを目指す．また，これらの対話コーパスを利用して，ALS を自動評価するための手法の開発などに取り組むことを検討している．

謝辞

本研究は，令和 6 年度金沢大学と北陸先端科学技術大学院大学による「融合科学共同専攻」における分野融合型研究支援を受けて行われた．また，金沢大学の吉村晋平氏からカウンセリング会話データに関する助言を頂いた．

参考文献

[1] Charles J. Stewart, Jr. Cash, W.B.: Interviewing: Principles and Practices. W.C. Brown Company (1982)

[2] Chaszczewicz, A., Shah, R., Louie, R., Arnow, B., Kraut, R., Yang, D.: Multi-level feedback generation with large language models for empowering novice peer counselors. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4130–4161. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.227, https://aclanthology.org/2024.acl-long.227

[3] Chatzinikola, M.E.: Active listening as a basic skill of efficient communication between teachers and parents: An empirical study. European Journal of Education and Pedagogy **2**(6), 8–12 (Nov 2021). https://doi.org/10.24018/ejedu.2021.2.6.186, https://www.ej-edu.org/index.php/ejedu/article/view/186

[4] Demasi, O., Li, Y., Yu, Z.: A multi-persona chatbot for hotline counselor training. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3623–3636. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.324, https://aclanthology.org/2020.findings-emnlp.324

[5] Doas, M.: Are we losing the art of actively listening to our patients? connecting the art of active listening with emotionally competent behaviors. Open Journal of Nursing **05**, 566–570 (01 2015). https://doi.org/10.4236/ojn.2015.56060

[6] Eitel, K.E.: The effects of mindfulness and distress disclosure on emotional expression (2014), https://api.semanticscholar.org/CorpusID:141796461

[7] Gollapalli, S., Ang, B., Ng, S.K.: Identifying Early Maladaptive Schemas from mental health question texts. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 11832–11843. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.792, https://aclanthology.org/2023.findings-emnlp.792

[8] Harry Weger Jr., Gina Castle Bell, E.M.M., Robinson, M.C.: The relative effectiveness of active listening in initial interactions. International Journal of Listening **28**(1), 13–31 (2014). https://doi.org/10.1080/10904018.2013.813234

[9] Hoste, E., Kashani, K., Gibney, N., Wilson, F., Ronco, C., Goldstein, S., Kellum, J., Bagshaw, S.: Impact of electronic-alerting of acute kidney injury: Workgroup statements from the 15 adqi consensus conference. Canadian Journal of Kidney Health and Disease **3** (12 2016). https://doi.org/10.1186/s40697-016-0101-1

[10] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

[11] Longbrake, E.: The effects of ethnicity and characteristics of practitioners on disclosure of sexually sensitive information. The Ohio State University (2011)

[12] Misato Hirota, Rie Chiba, S.A.Y.H.K.I.C.G.H.F.K.Y.T.H.: Individual nurse-led active listening intervention for spouses of individuals with depression: A pre-/posttest pilot study. vol. 61, pp. 19–25. J Psychosoc Nurs Ment Health Serv. (2023). https://doi.org/10.3928/02793695-20230524-01

[13] Na, H.: CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2930–2940. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.261

[14] OECD: Health at a glance 2023: Oecd indicators. OECD Publishing (2023)

[15] Robert Elliott, Arthur C Bohart, J.C.W.D.M.: Therapist empathy and client outcome: An updated meta-analysis. Psychotherapy (Chic) **55**(4), 399–410 (Dec 2018). https://doi.org/10.1037/pst0000175

[16] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[17] Xiao, Z., Zhou, M.X., Chen, W., Yang, H., Chi, C.: If i hear you correctly: Building and evaluating interview chatbots with active listening skills. p. 1–14. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376131, https://doi.org/10.1145/3313831.3376131

[18] Zhang, C., Li, R., Tan, M., Yang, M., Zhu, J., Yang, D., Zhao, J., Ye, G., Li, C., Hu, X.: CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 13947–13966. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.830, https://aclanthology.org/2024.findings-acl.830

[19] 神村栄一, 小林奈穂美, 鈴.: DVD で学ぶ新しい認知行動療法 うつ病の復職支援. 星屑倶楽部 / 中島映像教材出版 (2011)