KokoroChat: 講習を受けたカウンセラーによるロールプレイを通じて収集された日本語カウンセリング対話データセット

斉 志揚 [1]　金子 拓正 [1]　高溝 恵子 [2,3,4]　浮世 満理子 [2,3,4]　稲葉 通将 [1,4]
[1] 電気通信大学　[2] 一般社団法人 全国心理業連合会
[3] 株式会社アイディアヒューマンサポートサービス
[4] 株式会社ラポールテクノロジーズ
{qizhiyang,t-kaneko,m-inaba}@uec.ac.jp

概要

言語モデルによる心理カウンセリング応答の生成には，高品質なデータセットが不可欠である．しかし，クラウドソーシングによる収集には作業者の教育が必要であり，実際のカウンセリング記録はプライバシーの問題から利用が難しい．近年は大規模言語モデル（Large Language Model; LLM）による対話データの自動生成も行われているが，多様性やカウンセリングプロセスの信頼性に欠ける場合が多い．本研究では，講習を受けたカウンセラー同士によるロールプレイにより，高品質な日本語の心理カウンセリング対話を収集した．その結果，6,589 件の長文対話とクライアントによる包括的なフィードバックを含むデータセット KokoroChat[1)] を構築した．このデータを用いた LLM のファインチューニングにより，カウンセリング応答生成において性能の向上が確認された．

1　はじめに

メンタルヘルスの不調に苦しむ人々は多く，世界的に重要な課題となっている [1]．しかし，限られた医療リソースのため，多くの人々が専門的な心理カウンセリングを受けられない状況にある [2]．この課題に対応するため，共感的な応答を生成する言語モデルの研究が進められており，高品質なデータセットの構築が重要視されている．例えば，Liu らはクラウドワーカーに専門的なサポートスキルを教育して ESConv データセットを構築し [3]，Li らは実際のクライアントと専門カウンセラーの対話を収集し，Client-Reactions データセットを作成した [4]．

しかし，心理カウンセリングは高度に専門的なコミュニケーションであり [5]，非専門家であるクラウドワーカーを教育するには高いコストと時間を要する．一方，実際のカウンセリング対話データの収集には，プライバシーや倫理面の制約が伴い，公開が困難である．そのため，これらの手動収集に基づく手法には実用上の限界が存在する．

近年，LLM は自然言語生成において大きな進展を遂げ，心理カウンセリング応答生成においても高い可能性を示している [6]．これを受け，LLM を活用して心理カウンセリング対話データを自動合成する研究が数多く行われている [7, 8, 9, 10, 11]．しかし，このように合成されたデータセットは均質になりやすく，対話の多様性に欠けるという課題が指摘されている [12]．さらに，表 1 に示すように，合成データセットは人手で収集された心理カウンセリング対話と比較して，対話あたりの発話数が著しく少なく，実際のカウンセリングが持つ対話の深みや現実性に課題を残している．

これらの課題に対処するため，本研究では，図 1 に示すように，ロールプレイによるデータ収集手法を採用する．本手法では，講習を受けたプロのカウンセラーおよび訓練生がカウンセラーとクライア

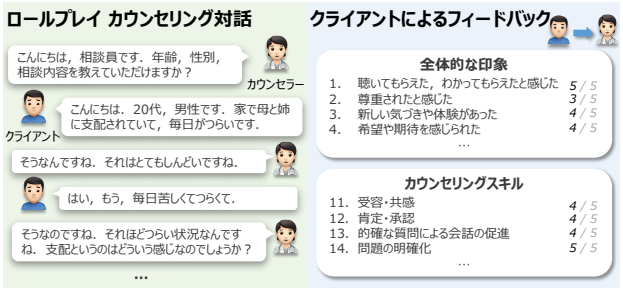

図 1　各 KokoroChat のデータは，ロールプレイの心理カウンセリング対話とクライアントからのフィードバックで構成されており，両役割はいずれも講習を受けたカウンセラーによって演じられている．

1）本データセットは https://github.com/UEC-InabaLab/KokoroChat にて公開している．

表 1　心理カウンセリング対話データセットの比較：LLM 合成（上段）と人手収集（下段）

データセット	人手作成	評価有無	評価項目数	言語	対話数	平均発話数
HealMe [8]	✗	✗	-	英語	1,300	6.0
ESD-CoT [9]	✗	✗	-	英語	1,708	23.4
CACTUS [10]	✗	✗	-	英語	31,577	31.5
SMILECHAT [11]	✗	✗	-	中国語	55,165	33.2
AUGESC [7]	✗	✗	-	英語	65,000	26.7
Anno-MI [13]	✓	✗	-	英語	133	72.9
ESConv [3]	✓	✓	2	英語	1,300	29.5
Client-Reactions [4]	✓	✓	4	中国語	2,382	78.5
KokoroChat	✓	✓	**20**	**日本語**	**6,589**	**91.2**

ントの役割を担い，テキストベースの SNS カウンセリングをシミュレーションする．本手法および本データセットには，以下の 4 つの特徴がある．(1) クラウドソーシング手法と比較して，参加者が専門性や講座の受講経験を有しているため，高い対話品質と専門性が確保されている．(2) 実際のカウンセリング対話を直接収集する方法とは異なり，ロールプレイに基づく対話であるため，プライバシーや倫理面のリスクを低減できる．(3) LLM によって自動生成されたデータと比べて，現実性の面で優位性がある．(4) 人手収集による心理カウンセリング対話データの中で最大規模であり，日本語データセットとして初めて公開されたものである．

2　関連研究

2.1　NLP における心理カウンセリング

近年，世界的なメンタルヘルスケアの需要拡大に伴い，自然言語処理（Natural Language Processing; NLP）分野では心理カウンセリングへの関心が高まっている．初期の研究はユーザの感情を理解し適切なフィードバックを行う共感的な応答生成に焦点を当てていた [14]．しかし，心理カウンセリングには共感だけでなく，ユーザの悩みを探り効果的なガイダンスを提供する能力も求められるため，Liu らは Emotional Support Conversation (ESC) タスクを提案した [3]．LLM の進化に伴い，Inaba らが行った研究では，GPT-4 のカウンセリング能力を専門家レベルと評価するなど，その応用可能性が注目されている [6]．ChatCounselor [15] や MeChat [11] といった特化型システムも登場しているが，これらは主に，クラウドソーシングによって収集された既存データや合成データを用いたファインチューニングに依存している．依然として高品質かつ多様な専門データの不足が課題であるため，本研究では講習を受けたカウンセラーによるロールプレイを通じて，高品質なデータセット KokoroChat を構築する．

2.2　カウンセリング対話データセット

言語モデルのカウンセリング能力向上にはデータセットが不可欠であり，主に人手による構築と LLM 合成の 2 種類が存在する．人手による構築の例として，動画から対話を抽出した Anno-MI [13]，クラウドワーカによる ESConv [3]，実際の相談ログに基づく Client-Reactions [4] がある．一方，LLM 合成はカウンセラーとクライアントの両方をシミュレーションする手法である．CBT に基づくプロンプトを用いた HealMe [8] や，多様な状況に応じて対話を生成する ESD-CoT [9] などが挙げられる．表 1 に示すように，LLM 合成データセットは規模こそ大きいものの，対話あたりの発話数が少なく，内容の多様性や均質性に課題が残る [12]．対照的に，本研究で提案する KokoroChat は，人手による収集としては最大規模でありながら，実際のカウンセリングに近い対話長と高い品質を兼ね備えている．

3　データ収集

高品質な心理カウンセリング対話データセットを構築するため，本研究では，講習を受けたプロのカウンセラーおよび訓練生がカウンセラーとクライアントの対話をシミュレーションするロールプレイ手法を採用し，専門性と現実性を確保した．

3.1　データ収集環境と手順

参加者のマッチング，ロールプレイ対話，およびクライアントフィードバックの収集を円滑に行うため，オンラインプラットフォームを開発した．参加者はこのプラットフォーム上で希望する役割を選択

し，自身の都合と役割の希望に基づいて対話のスケジュールを組むことができる．マッチング成立後，指定された時間にロールプレイ対話が行われる．この際，カウンセラー役はコンピュータの Web インターフェースを使用し，クライアント役はスマートフォンの LINE アプリを使用して対話を行う．この設定は，日本における実際の SNS 心理カウンセリングを反映したものである．

各対話セッションは通常 1 時間であるが，必要に応じて時間を調整することが可能である．対話開始前に，カウンセラー役が対話トピックやクライアントの心理状態を指定しない場合，クライアント役が自由に設定する．参加者のプライバシーを保護するため，すべてのクライアント役参加者に対し，実生活の悩みを話さないよう明示的に指示するとともに，個人を特定できる情報の共有を厳格に禁止している．セッション終了後，クライアント役はカウンセラー役のパフォーマンスを評価する．

3.2 参加者属性

データセットは 480 名の参加者で構成されており，その内訳は男性 117 名，女性 360 名，性別非公開 3 名である．参加者の年齢は 21 歳から 78 歳に及び，約 80%が 30 代から 50 代である．詳細な分布情報は付録 A に示す．参加者は全員日本語を母語としており，80%以上がクライアント役とカウンセラー役の両方を経験している．合計で 424 名がカウンセラー役を，463 名がクライアント役を担当した．

専門性および講習　参加者全員はオンライン心理カウンセリングに関する専門知識を有している．3 分の 1 以上は専門資格を持ち，実務経験がある．残りの参加者も，現時点では資格未取得であるが，専門資格を目指して 6 ヶ月から 1 年間の体系的な学習を行っている．さらに，全参加者が 10 時間の体系的なトレーニングプログラムを修了している．このプログラムには，オンラインテキストカウンセリングの特徴，カウンセラーの役割と倫理指針，および専門的なカウンセリングスキルと手順が含まれる．

3.3 クライアントフィードバック

各対話終了後，クライアント役はカウンセラー役のパフォーマンスを評価する．結果は即座にカウンセラー役に共有されるとともに，公平性と信頼性を確保するために管理者によって監視される．

フィードバックの項目は，国家資格である公認心理師の資格と博士号を持つ専門家の監修のもと，以下の 2 つの側面それぞれ 10 項目から構成される形で設計されている．

- **対話の全体的な印象**（例：理解・尊重，気づき・希望，協働性，対話の流暢さ，満足度）
- **カウンセリングスキルの評価**（例：共感・承認，質問スキル，要約力，問題・目標の明確化，行動促進，勇気づけと希望提示）

評価には 20 項目にわたる 6 段階のリッカート尺度（0〜5 点）を用い，合計スコアは最大 100 点である．評価項目の一覧は付録 B に示す．

表 2　KokoroChat の統計情報

項目	全体	カウンセラー	クライアント
対話数	6,589	-	-
話者数	480	424	463
総発話数	600,939	306,495	294,444
対話あたりの平均発話数	91.20	46.52	44.69
平均発話長	28.39	35.84	20.63

3.4 統計情報

本研究では，2020 年 3 月から 2024 年 9 月にかけて対話データを収集した．データ品質を確保するため，発話数が 30 未満のもの，対話時間が 30 分未満のもの，および 20 の評価項目すべてが「3」と評価されたもの（スコアの信頼性が低いため）を除外した．最終的なデータセットは 6,589 件の対話から構成される．統計的な詳細は表 2 に示す．

さらに，本データセットには 480 名の参加者が含まれており，その内訳はカウンセラー役が 424 名，クライアント役が 463 名である．これにより 4,900 通りの異なるカウンセラーとクライアントのペアが形成され，対話の多様性の確保に寄与している．

4 評価実験

本データセットの有用性を検証するため，心理カウンセリング応答生成タスクにおいて実験を行った．しかし，心理カウンセリングに特化した日本語データセットや日本語 LLM が不足しているため，他モデルとの直接比較は困難である．そこで本研究では，KokoroChat によるファインチューニングが，オープンソース LLM の心理カウンセリング性能の向上に寄与するかどうかを検証する．

対話データの前処理として，同一話者の連続発話を 1 つに結合した．モデルは，対話履歴 $D_t = \{u_1^C, u_2^S, u_3^C, \ldots, u_t^C\}$ を入力とし，次のカウンセ

ラー発話 u_{t+1}^{S} を生成する．ここで，u_i^C と u_j^S はそれぞれクライアント（C）およびカウンセラー（S）の発話を表す．

4.1 モデル

実験では，ベースモデルとして Llama 3.1 Swallow[2] を使用した．高品質なテストデータを確保するため，クライアントフィードバックのスコアが 99 または 100 である 118 件の対話をテストセットとし，残りをファインチューニングに用いた．

フィードバックのスコアは対話品質の指標となり得る．そこで，高評価対話と低評価対話を分離して学習を行い，以下の 3 条件でスコア差が応答生成性能に与える影響を検証した．

- **Kokoro-Low:** スコア < 70 の 3,870 対話（334,022 発話）で学習．[3]
- **Kokoro-High:** 70 ≤ スコア ≤ 98 の 2,601 対話（254,515 発話）で学習．
- **Kokoro-Full:** スコア ≤ 98 の 6,471 対話で学習．

また，商用 LLM との相対性能を把握するため，最先端モデルである GPT-4o [16] との比較を実施した．学習プロセスの詳細は付録 C に記載する．

4.2 自動評価

自動評価には，単語の重複度および多様性を測定する指標として，BLEU-n [17]，ROUGE-L [18]，Distinct-n [19] を用いた．結果は表 3 に示す．実験結果によると，Kokoro-High はほとんどの BLEU および ROUGE スコアで最も高い性能を示した．これは学習データの品質が高く，テストセットとの整合性が高いためと考えられる．対照的に，より大規模なデータセットを含む Kokoro-Full は，多様性の指標である Distinct において優れた性能を達成した．また，ファインチューニングを行っていないモデルに関しては，GPT-4o がすべての自動評価指標において Llama-3.1 を上回った．

4.3 人手評価

各モデルにより，ランダムに抽出した 100 件の異なる対話履歴に基づく応答を生成し，それらについて 5 名のプロカウンセラーによる独立した人手評価を実施した．評価にはペアワイズ比較を用い，カウンセラーはどちらの応答がより適切か（Win，Lose，Tie）を判定した．最終結果は多数決に従い，過半数が一致した場合はそれを採用し，それ以外または Tie が過半数の場合は Tie とした．

2） Llama-3.1-Swallow-8B-Instruct-v0.3 を使用した．
3） スコアの低い対話は発話数が少ない傾向があるため，データ分割のバランスを考慮し，閾値を 70 に設定した．

表 3 モデルの性能比較

モデル	B-1	B-2	B-3	B-4	R-1	R-2	R-L	D-1	D-2
Llama-3.1	17.32	9.13	4.77	2.25	23.81	7.37	16.96	1.04	6.86
GPT-4o	21.77	11.72	6.32	3.17	28.67	9.19	19.82	1.19	6.90
Kokoro-Low	25.39	15.30	8.69	5.39	33.38	14.05	27.28	2.42	12.98
Kokoro-High	**27.03**	**16.45**	**9.57**	**6.00**	**34.64**	**14.72**	28.00	2.33	13.08
Kokoro-Full	25.69	15.65	9.23	5.83	34.02	14.60	**28.10**	**2.48**	**13.24**

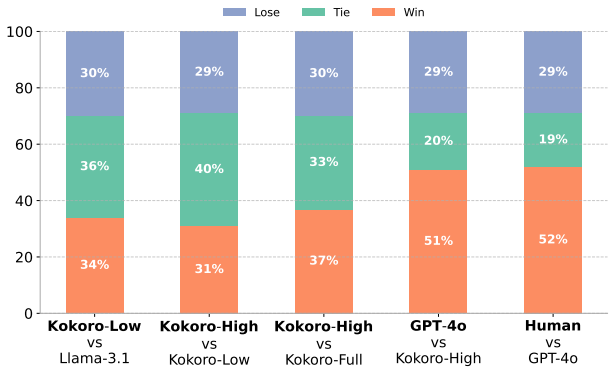

図 2 人手評価の結果．優勢モデルは**太字**で示す．

評価結果を図 2 に示す．Kokoro-Low と Llama-3.1 の比較から，KokoroChat の低スコア対話のみを使用した場合でも，オープンソース LLM の心理カウンセリング応答生成能力が向上することが示された．

また，自動評価の結果と同様に，データ量が少ないにもかかわらず，Kokoro-High は Kokoro-Low および Kokoro-Full のいずれよりも高い性能を示した．この結果は，モデル性能の向上において，高品質な学習データが重要であることを示唆している．しかし，モデル規模の差が極めて大きいことから，ファインチューニングしたモデルであっても，依然として GPT-4o には及ばない．一方で，GPT-4o の応答も，高評価を受けた人間のカウンセラーの応答と比較すると依然として明確な差が見られ，KokoroChat の品質の高さが裏付けられた．

5 おわりに

本研究では，ロールプレイ手法を用いて構築された，現時点で最大規模の人手収集による心理カウンセリング対話データセット KokoroChat を提案した．実験の結果，KokoroChat は心理カウンセリング応答生成における LLM の性能向上に寄与することが示された．

謝辞

本研究のクライアントフィードバック項目の設計にあたり，京都大学の杉原保史教授より貴重な助言を賜りましたことに，深く感謝申し上げます．

本研究は，JSPS 科研費 25H01382 の助成を受けたものです．

参考文献

[1] WHO. World mental health report: Transforming mental health for all, June 2022.

[2] SAMHSA. Behavioral health trends in the united states: Results from the 2014 national survey on drug use and health. Annual report, Substance Abuse and Mental Health Services Administration, September 2015.

[3] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In **ACL 2021**, pp. 3469–3483, Online, August 2021. Association for Computational Linguistics.

[4] Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. Understanding client reactions in online mental health counseling. In **ACL 2023**, pp. 10358–10376, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 463–476, 08 2016.

[6] Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. In **The 14th International Workshop on Spoken Dialogue Systems Technology**, 2024.

[7] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. AugESC: Dialogue augmentation with large language models for emotional support conversation. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1552–1568, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In **ACL 2024**, pp. 1707–1725, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In **ACL 2024**, pp. 13395–13412, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 14245–14274, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 615–636, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. Self-chats from large language models make small emotional support chatbot better. In **ACL 2024**, pp. 11325–11345, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In **ICASSP 2022**, pp. 6177–6181, 2022.

[14] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In **ACL 2019**, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support, 2023.

[16] OpenAI. Gpt-4o system card, 2024.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL 2002**, p. 311–318, USA, 2002. Association for Computational Linguistics.

[18] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[19] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **NAACL 2016**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[20] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[21] 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 **ACL 2025**, pp. 12424–12443, Vienna, Austria, July 2025. Association for Computational Linguistics.

A　参加者属性：年齢層と性別

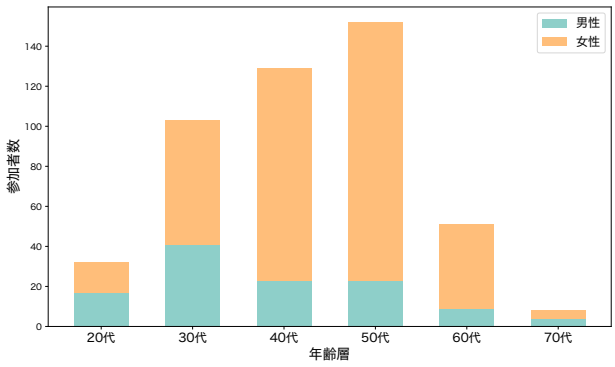

図 3　参加者の年齢および性別の分布

B　フィードバック評価項目

表 4　20 項目のクライアントフィードバック（各項目は 0〜5 の 6 段階評価）

カテゴリ	フィードバック項目
対話の全体的な印象	1. 聴いてもらえた・わかってもらえたと感じた
	2. 尊重されたと感じた
	3. 新しい気づきや体験があった
	4. 希望や期待を感じられた
	5. 取り組みたかったことを扱えた
	6. 一緒に考えながら取り組めた
	7. やりとりのリズムがあっていた
	8. 居心地のよいやりとりだった
	9. 全体として適切でよかった
	10. 今回の相談は価値があった
カウンセリングスキルの評価	11. 相談開始の円滑さ
	12. 相談終了のタイミング・円滑さ
	13. 受容と共感が示されていた
	14. 承認と肯定が得られた
	15. 対話を促進する効果的な質問があった
	16. 要点が効果的にまとめられていた
	17. 問題点が明確に整理された
	18. 会話の目標設定ができた
	19. 実行可能な提案があった
	20. 励ましと希望が与えられた

C　学習設定詳細

ファインチューニング段階　本研究では，QLoRA (Quantized Low-Rank Adaptation) [20] を用いて，4 ビット量子化された LLM に対する効率的なファインチューニングを行った．これにより，計算コストを抑えつつ高い性能を維持することを可能にした．

データセットの分割については，高品質な評価基準を確保するため，スコアが 99 または 100 の対話 118 件をテストセットとして選定した．残りのデータは，90%を学習用，10%を検証用に分割した．

ハイパーパラメータのチューニングにはグリッドサーチを用い，最適な構成を決定した．サーチは，optimizer，warmup steps，学習率の 3 つの主要なパラメータを対象とした．optimizer の候補には `adamw_torch_fused`，`adamw_8bit`，`paged_adamw_8bit` が含まれる．warmup steps は `100`，`300`，`500` で検証し，学習率は `1e-3`，`5e-4`，`2e-4`，`1e-4`，`5e-5` から選択した．評価結果に基づき，最終構成として optimizer に `adamw_8bit`，warmup steps に `100`，学習率に `1e-3` を採用した．学習は 4 枚の A100 40GB GPU を使用し，バッチサイズ 8，5 エポックで実施した．検証は `400` ステップごとに実施し，損失が最も低いモデルを最終モデルとして選択した．

推論段階　推論段階においても，モデルの性能を維持しつつ計算効率を最適化するため，4 ビット量子化を採用した．さらに，決定論的な出力を保証し，サンプリングのばらつきを排除して応答の一貫性を高めるため，`do_sample = False` および `temperature = None` を設定した．

さらに，本データセットを用いて，クライアントフィードバックスコアの予測タスクについても実験を行った．本稿では詳細な議論は割愛するが，本タスクの設定および結果については別途報告している論文 [21] に記載している．