対話における心情記述: 自然言語による機微かつ複雑な心情理解のためのベンチマーク

田中 義規 [1]　上原 隆一 [1]　井上 昂治 [2]　稲葉 通将 [1]
[1] 電気通信大学　[2] 京都大学
{y-tanaka,r-uehara,m-inaba}@uec.ac.jp
inoue.koji.3x@kyoto-u.ac.jp

概要

対話における感情認識の研究は，従来，離散的な感情カテゴリや連続的な次元で話者の感情を表現してきたが，複雑で機微な感情を十分に捉えることは困難である．本研究では，発話の背後にある話者の心情を自然言語で記述する新たなタスク「対話における心情記述 (Emotion Transcription in Conversation: ETC)」を提案する．本タスクのため，我々は話者の心情文付きの日本語対話データセット[1)]を構築した．さらに，ベースラインモデルの構築と評価を通じて，本タスクの実現可能性を示す．本データセットで fine-tuning を行うことで予測性能は向上したが，依然としてスコアは低く，ETC タスクの難しさが明らかとなった．

1　はじめに

対話における感情認識 (Emotion Recognition in Conversation; ERC) は，対話中の特定の発話における話者の感情を識別するタスクであり，人間と機械の自然で円滑なコミュニケーションを実現するために重要である．特に，共感的で人間らしい対話システムの構築には，システムがユーザの感情を正確に理解することが不可欠である．この分野の研究は過去 10 年間で大きく発展しており，深層学習や大規模言語モデル (LLM) の登場により，認識精度は実用レベルに近づいてきている [1].

また，現在までに数多くのベンチマークデータセットも構築されている [2, 3, 4]．これらは主に，離散的な感情カテゴリや感情次元にもとづくアノテーションを採用している．これらのアプローチは，データセットの定量的な分析や予測モデルの評価を容易にするという利点を有するものの，人間の複雑で機微な感情を十分に捉えることは難しい．

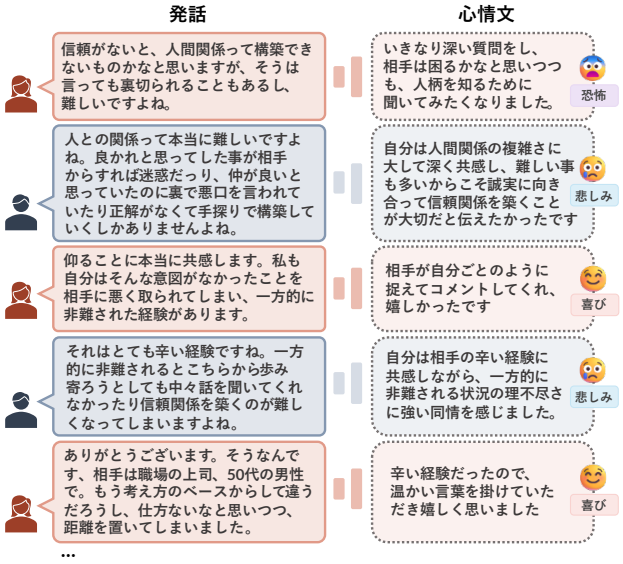

図 1　本研究で構築したデータセットからの対話および心情文の例．各心情文には，感情カテゴリもマルチラベルアノテーションされている．

そこで，本研究では感情認識のための新たなタスクとして，「対話における心情記述 (Emotion Transcription in Conversation; ETC)」を提案する．従来のアプローチと異なり，本タスクは，話者の心情を自然言語で記述することに焦点を当てる．我々の提案するアプローチは，感情カテゴリや数値的な枠組みでは捉えきれない，より機微で複雑な心情を表現できることが期待される．本研究で提案する ETC タスクを実現するための足掛かりとして，本研究では，発話に対応する心情文が付与された対話データセットの収集とベンチマークの確立に取り組む．まず，クラウドソーシングを通じて，日常的シナリオを想定した日本語の心情文付きテキスト対話データを収集する（図 1）．次に，収集したデータセットを用いて，対話文脈からこれらの心情文を生成するモデルを開発および評価することで，ETC タスクの実

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

現可能性を実証し，本タスクの難しさと今後の研究の方向性について議論する．

2 関連研究

近年 ERC は，人間らしい対話システムの構築や意見マイニングなどを目的として，自然言語処理分野において活発に研究されている．既存の研究では，高性能なモデルの開発に加え，ERC 研究を促進するためのベンチマークデータセットの構築も進められてきた．TV シリーズ「Friends」を元にした MELD は，多人数対話を収録したデータセットであり，感情の認識や遷移の研究のために広く活用されている [4]．また各発話には，Ekman の 6 基本感情 [5] と中立に基づく感情ラベルが付与されている．IEMOCAP は，台本に基づく対話と即興対話の両方を収録したデータセットであり [2]，マルチモーダルデータと感情ラベルを含む．DailyDialog は，人間による自然な二者間対話を収録したデータセットであり，Ekman の基本感情に加えて意図ラベルが付与されている [3]．EmotionLines は「Friends」を元にしたテキスト対話から構成され，MELD の前身となるデータセットである [6]．EmoryNLP も同様に「Friends」が元となっており，独自のデータ分割とアノテーションを採用している [7]．EmoContext は，3 ターンの短い対話に焦点を当て，簡略化された感情カテゴリを用いている [8]．認知的評価の理論に立脚した中国語データセットである CAPE は，LLM における感情の生起過程に注目したものである [9]．

ERC のためのデータセットに見られるように，従来の研究では，主に感情カテゴリや感情次元を用いて話者の感情を表現してきた．しかし，これらの枠組みは，人間の複雑で機微な感情を十分に捉えることは困難である．そこで本研究では，話者の心情を自然言語で記述するアプローチに着目し，心情文付きの対話データセットを構築する．本研究と同様に，一部の研究では話者の心理状態を自然言語で明示的に表現しようとしているが，こうした取り組みは主に LLM を用いて人工的に生成されたデータに依存している [10]．我々は，人間同士の対話からなるデータセットを構築する．

3 データセット構築

本研究では，ETC タスクのベンチマークとして，各発話に心情文と感情ラベルが付与された日本語対話データセットを構築した．オンライン上で，テキスト形式の対話データを収集した．

3.1 対話の収集

本研究では，クラウドソーシングプラットフォームであるクラウドワークス[2)]上で参加者を募集した．対話の収集に先んじて，話者の属性情報の把握や分析を可能にするため，参加者の Big Five 性格特性を計測した．性格特性の計測には，TIPI-J（日本語版 Ten-Item Personality Inventory） [11] に基づく 10 項目の質問票を用いた．

我々は，話者の豊かで自然な感情表現を引き出すため，EmpatheticDialogues の対話設定 [12] を採用した．本設定ではまず，参加者は「スピーカー」と「リスナー」のいずれかの役割を割り当てられる．また，対話を実施するペアには特定の感情ラベルが与えられる．スピーカーはその感情を感じた具体的な体験をリスナーに語り，リスナーはスピーカーの話に反応することが求められた．各対話はスピーカーの発話から始まり，交互に 5 ターン，合計 10 発話で終了するものとした．

対話の際，参加者は，発話を入力した直後に発話時点での心情を記述するよう求められた．本データ収集においては，心情を「対話の参加者がその発話時に抱いていた内面的な感情状態や意図を言語化したもの」と定義した．また，発話および心情文は 5 文字以上であることを条件とし，AI の使用および個人情報の入力や誹謗中傷は禁止とした．加えて我々は，すべてのデータに対して品質チェックを実施し，個人情報や倫理的に不適切な内容が含まれるデータがないかを確認した．

3.2 データセットの統計量

本研究で構築したデータセットは，合計 1,002 件の対話から構成される．収集されたデータの例を図 1 に示す．本データセット構築には，199 名のクラウドワーカが参加した．ワーカの参加した対話回数の中央値は 6.0，最大値は 38 であった．我々は，32 の感情ラベルに対して均等に対話を収集することを目指した．結果として，各感情ラベルに対する対話数は最小 30 件，最大 32 件となった．また，収集したデータセットの主な統計情報を表 1 に示す．スピーカーの発話の長さはリスナーの発話よりもやや長い傾向が見られるが，心情文の長さは両者で同等であった．

2） https://crowdworks.jp

対話数	1,002
発話数 / 心情文数	10,020
発話の平均長（文字数）	42.72
スピーカー	44.64
リスナー	40.80
平均心情文長（文字数）	28.89
スピーカー	28.92
リスナー	28.86

表 1　ETC データセットの統計情報

3.3　感情ラベルのアノテーション

収集した心情文は，話者の複雑な心情を表現するものの，それらを定量的に分析することが困難である．そこで本研究では，収集した各心情文に感情ラベルを追加でアノテーションした．これにより，発話の背後にある感情の定量的分析だけでなく，従来の ERC タスクへの適用が可能となる．我々は，従来研究 [6, 4, 3] で広く用いられている Ekman の 6 基本感情 [5]（喜び，悲しみ，恐怖，怒り，驚き，嫌悪）に，「該当なし」を加えた 7 カテゴリを採用した．また，複数の感情を含む心情文に対応するため，マルチラベル方式を採用し，いずれにも該当しない場合は「該当なし」のみを付与した．各心情文は，クラウドワークス上の異なる 3 人のアノテータによりアノテーションされ，先行研究 [2, 6] に従い多数決により最終ラベルを決定した．アノテーションされた感情ラベルの分布，およびアノテータ間の一致度は付録 A に記載する．

4　実験

本研究で構築したデータセットが ETC タスクに有用であるかを評価するため，我々はベースラインモデルを構築し，その性能を評価した．実験にあたり，データセットを訓練，検証，テストセットへ 8:1:1 の割合で分割した．

4.1　タスク定義

ETC タスクでは，話者自身が記述した心情文を予測することが求められる．本タスクにおいて，モデル $\mathcal{M}$ には，n 番目の発話までの対話文脈 $C_n = \{(u_1, s_1), (u_2, s_2), \ldots, (u_n, s_n)\}$ が与えられる．u_i は i 番目の発話，s_i はその話者を表す．本タスクの目的は，C_n にもとづいて話者 s_n による発話 u_n の心情文 e_n を予測することである．すなわち，$e_n = \mathcal{M}(C_n)$ である．

4.2　モデル

本研究では，実験時点での最新の対話モデルである **GPT-4.1**[3] と，日本語に特化したオープンソースの LLM である **Llama-3.1-Swallow**[4] [13, 14] の予測性能を調査する．両モデルとも，zero-shot 学習と 4-shot 学習による性能を評価した．加えて，本研究で構築したデータセットで fine-tuning した Llama-3.1 の性能も評価した．このモデルでは，結果の頑健性を考慮し，5 つの異なる乱数シードで学習したモデルの平均性能を報告する．詳細な学習設定は付録 B に記載する．

4.3　評価指標

4.3.1　従来型自動評価指標

生成された心情文の品質を評価するため，語彙の重なりを測る指標である BLEU [15] と ROUGE [16]，および意味的類似度を測る指標である BERTScore [17] の 3 つの従来型自動評価指標を採用した．

4.3.2　細粒度で解釈可能な忠実性評価指標

近年，LLM を評価器として用い，人間による判断に近い結果を示す自動評価手法が提案されている [18]．しかし，「自分の感情が理解されないことへの怒りと同時に，悲しみも感じている」のような複数の要素を含む心情文に対し，単一の評価値で整合性を測ることは困難である．この課題に対処するため，我々は FActScore [19] に着想を得た，内容の忠実性をきめ細かく評価する手法を用いる．この手法は以下の 2 段階の処理から構成される．

心情文の分割．まず，予測心情文と正解心情文のどちらか一方の心情文を，それぞれが単一の情報を含む情報単位に分割する．例えば，「自分の感情が理解されないことへの怒りと同時に，悲しみも感じている」は，「私は私の気持ちが理解されないことに怒っている」「私は私の気持ちが理解されないことが悲しい」という 2 つの情報単位に分割される．

情報単位の支持判定．次に，分割後の各情報単位が，もう一方の心情文によって支持されるかを，支持/不支持/中立の 3 クラスに分類する．例えば，情報単位「私は私の気持ちが理解されないことに怒っている」は，心情文「自分の感情が理解されないことへの怒りと同時に，悲しみも感じている」に含ま

3）`gpt-4.1-2025-04-14`
4）Llama-3.1-Swallow-8B-Instruct-v0.3

Models	Setting	B-1	B-2	B-3	B-4	R-1	R-2	R-L	BS	Prec.	Rec.	F1	# Units (SD)
GPT-4.1	zero-shot	16.89	7.99	3.93	2.09	23.61	4.87	17.93	57.66	14.73	**42.27**	13.99	3.39 (1.14)
	4-shot	26.89	13.34	7.36	4.12	28.06	5.78	22.98	59.67	20.59	27.40	13.78	1.96 (0.76)
Llama-3.1	zero-shot	17.40	7.20	3.55	1.84	20.25	3.08	16.26	55.24	9.18	17.98	5.77	1.99 (1.04)
	4-shot	29.41	14.86	8.46	4.84	27.51	5.58	23.22	59.95	14.83	14.18	7.84	1.58 (0.58)
	fine-tuning	**36.07**	**23.01**	**15.59**	**9.98**	**31.95**	**8.79**	**28.54**	**62.64**	**28.50**	19.71	**14.29**	1.39 (0.60)
Reference	–	–	–	–	–	–	–	–	–	–	–	–	1.35 (0.64)

表 2 ETC タスクの自動評価結果．BLEU (B), ROUGE (R), BERTScore (BS), Precision (Prec.), Recall (Rec.), F1-score (F1) および心情文に含まれる情報単位数 (# Units) を報告する．性能の最良値は**太字**，次点は下線で示されている．

れるため，「支持」と判定される．「中立」は，主要素は支持されるものの，感情の理由のような副次的要素の真偽が判断できない場合に該当する．

我々は，上述の両ステップの実装に Gemini-2.5-Flash[5] を使用した．その後，各予測心情文に対して以下のスコアを算出する．

Precision．予測心情文を情報単位に分割し，それらのうち正解文によって支持された情報単位の割合．

Recall．正解文を情報単位に分割し，それらのうち予測された心情文に含まれる情報単位の割合．

F1-score．Precision と Recall の調和平均．

なお，「中立」と分類された情報単位は正解としてカウントしない．また，正解文に心情を表す記述が存在しないときは上記のスコアが定義できないため，本研究では評価対象から除外した．評価対象のデータであっても，予測文に心情にあたる記述が存在しない場合には，上記スコアをすべて 0 とした．

4.4 結果と考察

表 2 に ETC タスクの自動評価結果を示す．BLEU, ROUGE，BERTScore といった従来指標において，ETC データセットで fine-tuning された Llama-3.1 が最も高い性能を示した．また，4-shot のモデルが zero-shot のモデルを一貫して上回るという傾向も確認された．細粒度評価では，Precision はこれらの従来指標と同様の傾向を示す一方，Recall では zero-shot の GPT-4.1 が他モデルを大幅に上回る結果となった．総合的な性能指標である F1 スコアでは，fine-tuning された Llama-3.1 が最良であり，GPT-4.1 も競争力の高い結果を示した．

細粒度評価指標での結果を，心情文に含まれる情報単位数（表 2 の # Units 列）の観点から考察する．Precision において最も高いスコアを達成した Llama-3.1 (fine-tuning) は，正解文とほぼ同程度の情報単位を含む心情文を生成する一方，zero-shot 設定のモデル，特に GPT 系モデルは 1 つの心情文あたりにより多くの情報単位を含む傾向がある．このような心情文は，正解文に含まれる情報を広範にカバーする可能性が高まるため Recall の向上につながるが，正解文に含まれない冗長な情報により Precision は低下する．対照的に，Llama-3.1 (fine-tuning) は他のモデルよりも少ない情報単位で最も高い Precision を達成しており，fine-tuning が心情文の正確な予測に有効であることを示唆している．

実験結果から，話者の心情文を予測することの難しさが浮き彫りとなった．F1 スコアは最良のモデルでさえ 14.29%に留まっており，Precision と Recall の間に大きな不均衡も見られる．これらの結果は，ETC タスクが現状のモデルにとって挑戦的な課題であり，本データセットが，心情理解に関する今後の研究を進める上で重要な意義を持つベンチマークとなることを示唆している．改善の方向性として，chain-of-thought のようなプロンプト設計の工夫や，RLHF や対照学習などの高度な fine-tuning 手法の探究，また，本研究で収集した性格特性の活用などが考えられる．性格特性は，先行研究において対話における心理状態との関連性が示唆されており [10]，有効な手がかりとなりうる．

5 おわりに

本研究では，話者の複雑で機微な心情を捉えた心情文を予測する，対話における心情記述 (ETC) という新たなタスクを提案した．また，ETC タスクのための日本語対話データセットを構築した．実験では，データセットによる fine-tuning が ETC タスクの性能向上に有効である一方，依然としてその性能には課題が残っていることも明らかとなった．

5) https://ai.google.dev/gemini-api/docs/models

謝辞

本研究は、科研費学術変革領域研究（B）（25H01382）の支援を受けた。

参考文献

[1] Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. Deep emotion recognition in textual conversations: A survey. **Artificial Intelligence Review**, Vol. 58, No. 1, pp. 1–37, 2025.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, Vol. 42, No. 4, pp. 335–359, Oct 2008.

[3] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In **International Joint Conference on Natural Language Processing (IJCNLP)**, pp. 986–995, Taipei, Taiwan, Nov 2017. Asian Federation of Natural Language Processing.

[4] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Rishi Naik, Erik Cambria, and Alexander Hoffmann. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In **Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 527–536, Florence, Italy, Jul 2019. Association for Computational Linguistics.

[5] P. Ekman, W. V. Friesen, M. J. O'Sullivan, A. K. Chan, I. Diacoyanni-Tarlatzis, K. G. Heider, R. Krause, W. A. LeCompte, T. K. Pitcairn, P. E. Ricci-Bitti, K. R. Scherer, M. Tomita, and A. Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. Vol. 53, pp. 712–717, 1987.

[6] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. EmotionLines: An emotion corpus of multi-party conversations. In **International Conference on Language Resources and Evaluation (LREC)**, 2018.

[7] Seyyed M. Zahiri and Jinho D. Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In **AAAI Workshops**, 2018.

[8] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In **Proceedings of the 13th international workshop on semantic evaluation**, pp. 39–48, 2019.

[9] June M. Liu, He Cao, Renliang Sun, Rui Wang, Yu Li, and Jiaxing Zhang. CAPE: A Chinese dataset for appraisal-based emotional generation in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Findings**, pp. 6291–6309, 2025.

[10] Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. **AAAI Conference on Artificial Intelligence**, Vol. 39, No. 2, pp. 1520–1528, 2025.

[11] Atsushi Oshio, ABE Shingo, and Pino Cutrone. Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j). **Japanese Journal of Personality**, Vol. 21, No. 1, 2012.

[12] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In **First Conference on Language Modeling**, 2024.

[14] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **First Conference on Language Modeling**, 2024.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**, ACL '02, p. 311–318, USA, 2002. Association for Computational Linguistics.

[16] 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.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[19] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

対話文脈	**話者 A**: この間、夜に車を運転していたら無灯火の自転車が飛び出してきて…本当に怖い思いをしました。 **話者 B**: 想像するだけで怖いです。寿命が縮みそうな出来事でしたね。 **話者 A**: 心臓が止まりそうでしたよ、事故にならなくてほっとしましたが、あとから怒りが湧いてきましたね。
出力	**正解文**: 相手が心情を理解してくれたことが嬉しく、気持ちをもっときょうゆうしたくなりました。 **GPT-4.1 (zero-shot)**: 自分が本当に危険な目にあったことへの恐怖と、事故にならずに済んだ安堵感、そして無灯火で飛び出してきた自転車に対して後から怒りがこみ上げてきた気持ちが混ざっている。相手にもそのときの自分の動揺や怒りを理解してほしいと思いながら話している。 **GPT-4.1 (4-shot)**: 相手に自分の恐怖や怒りの気持ちを理解してもらいたかったです。 **Llama-3.1 (zero-shot)**: 本当に怖かったし、怒りも感じています。 **Llama-3.1 (4-shot)**: 相手に自分の恐怖と怒りを理解してもらいたいと思っています。 **Llama-3.1 (fine-tuned)**: 相手の共感に嬉しく思い、さらに詳しく話を聞いてもらいたいと思いました。

表 3　心情記述のケーススタディ．各出力は話者 A の最終発話に対する心情文である．

Emotion	Spk. (%)	Lsn. (%)	All (%)	κ
喜び	29.9	20.0	25.0	0.603
悲しみ	10.8	8.8	9.8	0.359
恐怖	7.1	5.3	6.2	0.476
怒り	3.7	2.8	3.2	0.529
驚き	2.4	4.7	3.5	0.560
嫌悪	6.3	3.3	4.8	0.233
該当なし	43.4	57.2	50.3	0.400
全体	–	–	–	0.533

表 4　心情文に付与された感情ラベルの出現率とアノテータ間の Fleiss' kappa．各列名は対話における役割を示す: スピーカー (Spk.)，リスナー (Lsn.)，および全体 (All).

A　感情ラベルの統計量

表 4 に，心情文に付与された感情ラベルの出現率とアノテータ間の一致度を示す．アノテータ間の一致度は Fleiss の κ 係数を用いて算出した．全体として，50%近くの心情文は「該当なし」とラベル付けされている．感情ラベル別にみると，「喜び」が最も高い頻度で表現されている．役割別にみると，スピーカーの心情文では，リスナーのものよりも「該当なし」以外の感情ラベルの出現割合が高い．

B　実装の詳細

Llama-3.1-Swallow の教師あり fine-tuning は，2 基の NVIDIA A100 80GB GPU を用いて実行した．学習時のバッチサイズは 8，学習は 2 エポック実施した．また，メモリ効率化のため 4-bit 量子化を適用し，`PagedAdamW8bit` オプティマイザを使用し学習を行った．検証は 200 ステップごとに実施した．ハイパーパラメータのチューニングには，学習率を{1e-05，5e-05}，warm-up ステップ数を{300，700}としてグリッドサーチを実施した．結果として，学習率 1e-05，warm-up ステップ数 300 を選択した．推論時には，結果の再現性を考慮し，Llama-3.1 では `do_sample=False`，GPT-4.1 では `temperature` を 0.0 に設定した．

C　ケーススタディ

表 3 に，評価実験における各モデルによる心情文の生成例を示す．この例では，話者 A の最終発話には過去の出来事への安堵感と怒りが明示的に述べられている．一方，正解となる心情文には，対話相手の共感的な反応から生じた幸福感が表現されており，発話に表出されている感情との間にギャップが存在する．

こうした事例において，多くのモデルは話者の真の心情を捉えられていない．表 3 に示すように，GPT-4.1 および zero-shot，4-shot の Llama-3.1 モデルの出力は，発話に明示的に述べられた否定的な感情にのみ取り上げている．例えば，最も長い生成文を出力した GPT-4.1 の zero-shot 設定では，発話に表出している感情を記述するだけでなく，それらを拡張して対話相手の理解を求める内容となっているが，話者が実際に感じていた幸福感を予測することはできなかった．一方，本データセットで fine-tuning された Llama-3.1 モデルは，対話相手との共感的なやり取りから生じた話者の幸福感を予測している．このように発話で直接は表現されない心情を捉えることは，ETC タスクにおける重要な課題である．実際，fine-tuning されたモデルであっても予測に失敗するケースが確認されている．