

対話内難易度変化に追従する第二言語学習支援対話システム

森岡 拓¹ 高山 隼矢¹ 梶原 智之^{1,2}¹ 愛媛大学 ² 大阪大学

{morioka@ai., takayama@ai., kajiwara@}cs.ehime-u.ac.jp

概要

本研究では、対話トピックごとに変わる第二言語学習者の知識や言語能力に、対話の中で動的に追従する対話システムを構築する。先行研究では、第二言語学習者の会話能力向上の支援のために、インプット仮説における近接発達領域内に発話難易度を制御する対話システムが構築されてきた。しかし、対話トピックごとの言語能力差に起因するユーザの発話難易度の変化を考慮しておらず、変動する近接発話領域に対応する発話を生成できない。そこで対話システムが、対話の前半と後半で発話難易度が変化するユーザモデルに追従しているか、発話対の難易度差をもとに評価するフレームワークを提案する。LLMに最小限の指示を与えて会話させた場合と比較して、プロンプトを工夫したところ平均7.2%、選好ペアを作成してDPO学習したQwen3-14Bでは12.5%のスコア向上を達成した。

1 はじめに

第二言語学習者の会話練習機会の確保にはコストがかかる。英会話教室では会話相手となるチューターが1対1もしくは少人数クラスで対応する必要があり、留学には渡航費や滞在費を含めてコストがかかる。また、第二言語を実際に他人に対して発話することに不安を感じる学習者もいる[1]。近年普及している大規模言語モデル(LLM: Large Language Model)は対話形式のインタフェースを持っており、第二言語学習者の会話練習機会の確保と、対人対話の練習における不安の低減が見込める。

第二言語学習における学習者のインプットは、易しすぎず、難しすぎない近接発達領域内に収まるのが重要である[2]。そのため、これまでの第二言語学習者向け対話システムの開発では、所与の発話難易度を満たすように対話システムの発話を制御する手法が提案されてきた[3-5]。しかし、第二言語学習者の発話難易度は、学習者の話題への親しみ度

合いによって変動する[6,7]。これに対応して近接発達領域内の発話を提供するためには、対話中で動的にユーザの発話難易度に適応する必要があるが、LLMのこの能力を直接評価した研究は存在しない。

そこで、本研究ではトピックを対話の中で変化させ、それに応じて発話難易度が変化するユーザシミュレーションモデル(ユーザモデル)を導入する。このユーザモデルと対話システムの対話ログから、対話システムの発話難易度変化への適応能力を評価するフレームワークを提案する。対話ログから異なる会話トピックのターンの発話対を抽出し、難易度差を語彙と文法・文構造の2つの面から評価する。これはLLM同士を対話させ、そのログを評価するself-chat[4]と関連するが、ユーザモデルの発話難易度を前半と後半で変化させる点が異なる。

ユーザモデルの発話難易度変化に追従する能力を向上させる手法を2つ提案する。プロンプトによる発話難易度制御の手法と、DPO学習を利用する手法である。実験では、複数のLLMを利用して自動評価と人手評価で提案手法の有効性を検証した。自動評価においては、シンプルな対話指示のプロンプトに対して、プロンプトによる発話難易度制御手法とDPO手法の両方でスコア向上が確認できた。また、自動評価と人手評価のスコアの間で高いスパイマン順位相関と、主観的なタスクとしては妥当なCohen's Kappa係数が確認できた。

2 評価フレームワーク

第二言語学習者の対話トピックによる発話難易度差への対話システムの追従能力を測るために、擬似的なユーザモデルを使った評価フレームワークを提案する。概要を図1に示す。まず、難易度指示に対する追従性が高いLLMをユーザモデルとして選定する。また、対話トピック・ペルソナ・ユーザモデルに対する難易度指示からなるデータセットを構築する。マルチターン対話の前半と後半でユーザモデルの対話トピックと難易度を変化させて、評価対象

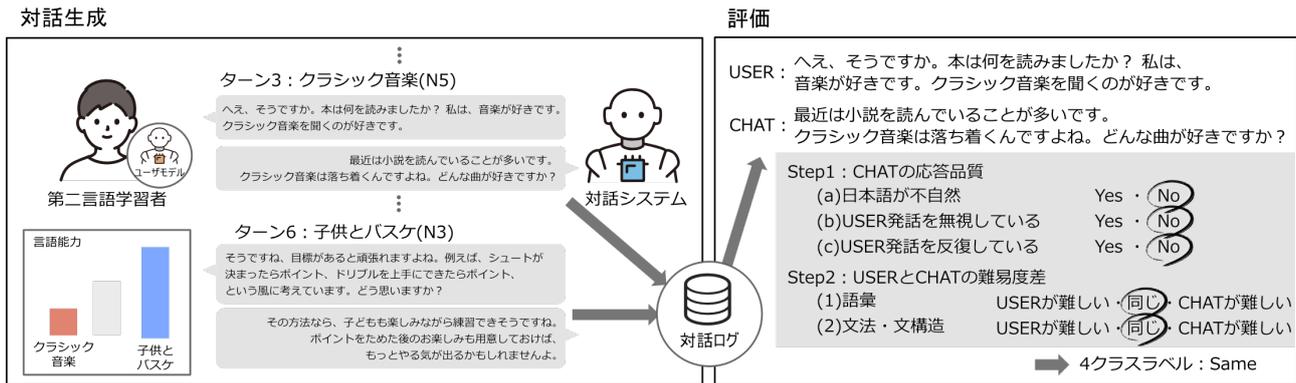


図1 評価フレームワーク概要

のモデルと対話させる。前半と後半の対話から、1ターンずつのユーザーモデル (USER) 発話と対話システム (CHAT) 発話を抽出し、以下の評価基準に従って評価する。

評価観点 難易度差評価に適する発話か step1 で確認し、step2 で難易度差を評価する。step1 では (a) 日本語が不自然である、(b) ユーザの発話を無視している、(c) ユーザの発話を単に反復している の3項目を Yes か No の二択で評価する。step2 では (1) 語彙、(2) 文法・文構造について2文を比較し、USER 発話の方が難しい (U) か、同じ難易度 (S) か、CHAT 発話の方が難しい (C) かの3択で評価する。

ラベル分類 5観点の結果を4クラスのラベルに分類する。**Reject** はCHAT 発話が難易度差評価に適さない場合を示す。step1 のいずれかが Yes の場合に割り当てる。**UserHarder** はUSER 発話がCHAT 発話より難しい場合を示す。語彙と文法・構文の2観点の両方がU、もしくはUとSの組み合わせのとき選択される。**Same** はUSER 発話とCHAT 発話の難易度差が小さい場合を示す。語彙と文法・構文の2観点の両方がSのとき、もしくはUとCの組み合わせで2観点で反対の傾向の場合に選択される。**ChatHarder** はCHAT 発話がUSER 発話より難しい場合を示す。語彙と文法・構文の2観点の両方がC、もしくはCとSの組み合わせのときに選択される。

3 提案手法

対話システムの難易度追従能力を向上させる4つの手法を提案する。プロンプトは図3に示す。

Baseline シンプルな対話指示のみをsystem プロンプトとして与え、対話履歴をassistant ロールとuser ロールで交互に与える。

Detailed Baseline に相手の発話難易度に揃えた発話をする指示を追加する。

JSON シングルターンプロンプトでDetailedと同様の指示をする。この手法は、LLMのチャットテンプレートで対話履歴を与えると、指示チューニング時の訓練データの対話応答の傾向が優先されプロンプト指示が効きづらいという仮定から、1つのプロンプトに対話履歴を全て含めて制御性を向上させる意図で導入した。1つのuser ロールの発話に、対話履歴をUSERとCHATの発話として埋め込み、JSON形式で出力するように指示を追加する。

DPO DPO [8] を利用して発話を制御する。学習で利用する選好ペアは複数モデルの出力から作成する。まず、学習データセットを用いて対象モデルとユーザーモデルの対話ログを作成する。この対話ログを利用して、他の候補モデルで評価対象ターンの発話候補を生成する。発話候補を自動評価して、step1 がすべてfalseで、step2 がすべてSameの発話を正例、候補の中で最もペナルティが大きい発話を負例とするペナルティは、step1のYesの数、step2のSame以外の数の総和で計算される。

4 実験

各提案手法と自動評価の妥当性を検証するために、評価フレームワークに従って、複数のLLMを対象に実験した。

4.1 実験設定

評価フレームワークの各部分の実験設定を述べる。実験ではLLMの推論にはvllmライブラリ [9]、その他のモデルの学習と推論にはHuggingFace Transformersライブラリ [10] を利用した。

ユーザーモデル選定 JPersonaChat¹⁾ [11] からペルソナと対話履歴を収集し、プロンプトで難易度

1) <https://github.com/nttctslab/japanese-dialog-transformers>

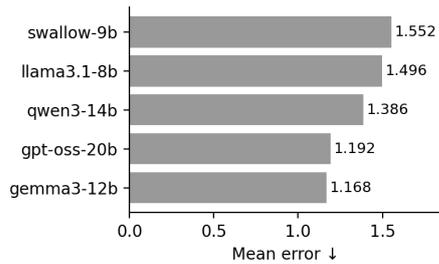


図2 ユーザモデル選定

指示を与えて LLM に発話を生成させる予備実験を行った。日本語能力試験 (JLPT) の、N1 から N5 に向かって順に平易になっていく合計 5 段階のラベルを利用した。200 件の対話履歴からそれぞれ 5 発話目までを抽出し、5 段階の難易度指示プロンプトを用意して、各モデルに合計 1,000 回発話を生成させた。比較モデルは、gpt-oss-20b²⁾ [12], Qwen3-14B³⁾ [13], gemma3-12b⁴⁾ [14], Swallow-9b⁵⁾ [15], Llama-3.1-8B⁶⁾ [16] である。miyataらの研究 [17] に従って、文対の難しい側を判定する分類器を訓練し、生成した発話をペアワイズ比較で並べ替えて難易度ラベルを付ける。分類器は日本語平易化パラレルデータセット MATCHA⁷⁾ [18] で ModernBERT⁸⁾ [19] をファインチューニングして構築する。MATCHA の 14,000 件を学習用、1,000 件を開発用と評価用として 5 エポック学習したところ、評価用データに対する accuracy が 0.93 となった。難しいと判断された回数の上位から 100 件ずつ N1 から順にラベルを割り当てる。割り当てたラベルと、指示したラベルの間の差の平均値を Mean error として図 2 に示す。gemma3-12b が最高の難易度制御能力であったので、ユーザーモデルとして採用する。

データセット構築 ユーザーモデルに与えるペルソナ・対話トピック・目標難易度のトリプルを作成する。ペルソナは JPersonaChat データセットから抽出した。GPT-5⁹⁾ の API で、ペルソナをもとに興味がありそうなトピックと話題リスト、興味がないようなトピックと話題リストをそれぞれ生成した。第二言語学習者向けの平易さと、ユーザーモデルのトピック

クによる難易度差の表現の容易さを考慮し、N5 と N3 の難易度ペアを採用する。興味がありそうなトピックにはより発話難易度が高い N3 を、興味がないようなトピックには発話難易度が低い N5 を割り当てた。前半と後半でどのトピックを出現させるかはランダムに決定した。学習用に 4,796 対話分、評価用に 100 対話分のデータセットを作成した。

対話ログの生成 ユーザーモデルと評価対象の対話システムの間で、6 ターンの対話を生成する。ダミー発話「こんにちは。最近どうですか?」から対話を開始する。ダミー発話はユーザーモデルの入力のみ利用される。対話トピックによる難易度変化を表現するために、作成したデータセットから前半 3 ターンと後半 3 ターンで異なるトピックと発話難易度をユーザーモデルに与える。4 ターン目にはユーザーモデルに難易度とトピックの転換を促すプロンプトを与える。また、4 ターン目以降のユーザーモデルには、3 ターン目以前の対話履歴を与えない。ユーザーモデルのペルソナは対話全体で共通である。

評価 評価には 3 ターン目と 6 ターン目のユーザーモデルと対話システムの発話対を用いた。5 つの評価観点について評価した後、4 ラベルに分類した。

アンサンブル 評価の安定性を高めるために、自動評価と人手評価の両方で、同じ発話ペアに対して 3 つ評価ラベルを付けた後にアンサンブルにより最終的なラベルを決定する。4 クラスラベルに対して多数決で最終ラベルを選択するが、ラベルが重複しなかった場合は Same ラベルを割り当てる。

手法の実装 Baseline, Detailed, JSON では LLM に対して各手法に対応するプロンプトを与えた。DPO 手法は Qwen3-14B 対象とする。Qwen3-14B の JSON 手法でベースとなる対話ログを生成した後、Baseline 手法で性能が低い gpt-oss-20b と Llama-3.1-8B で追加の発話候補を生成した。合計で 3 つの発話候補を収集した。これらを自動評価し、最終的に 4,726 件の選好ペアを作成した。trl ライブラリ¹⁰⁾を利用して LoRA [20] で学習した¹¹⁾。推論時は Baseline プロンプトを利用した。

自動評価 gpt-oss-20b, Qwen3-14B, gemma3-12b, Swallow-9b, Llama-3.1-8B, GPT-5.1¹²⁾ を対象に LLM-

2) <https://huggingface.co/openai/gpt-oss-20b>
 3) <https://huggingface.co/Qwen/Qwen3-14B>
 4) <https://huggingface.co/google/gemma-3-12b-it>
 5) <https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1>
 6) <https://huggingface.co/meta-llama/Llama-3.1-8B>
 7) <https://github.com/EhimeNLP/matcha>
 8) <https://huggingface.co/answerdotai/ModernBERT-base>
 9) <https://platform.openai.com/docs/models/gpt-5>, 利用期間: 2025/12/07 から 2026/01/04

10) <https://github.com/huggingface/trl>
 11) LoRA のパラメータは rank: 8, Alpha: 32, Dropout: 0.05, 学習パラメータは batch size: 2, epoch: 3, Beta: 0.1, 学習率: 5e-5, Alpha: 1.0, warmup ratio: 0.03, weight decay: 0.01, eval steps: 500 であり, eval loss が最小のモデルを選択した。
 12) <https://platform.openai.com/docs/models/gpt-5.1>, 利用期間: 2025/12/07 から 2026/01/04

表 1 各手法と難易度差の評価

手法	ターン	モデル	SameRate ↑	UserHarderRate ↓	ChatHarderRate ↓	RejectRate ↓	avg SameRate ↑
Baseline	マルチ	GPT-5.1	<u>0.125</u>	0.070	0.705	0.100	-
Baseline	マルチ	gpt-oss-20b	0.275	0.180	0.535	0.010	
Baseline	マルチ	Qwen3-14B	0.295	0.240	0.430	0.035	
Baseline	マルチ	gemma3-12b-it	0.280	0.430	0.180	0.110	0.281
Baseline	マルチ	Swallow-9b	<u>0.300</u>	0.395	0.250	0.055	
Baseline	マルチ	Llama-3.1-8B	0.255	0.310	0.315	0.120	
Detailed	マルチ	gpt-oss-20b	0.270	0.210	0.500	0.020	
Detailed	マルチ	Qwen3-14B	0.285	0.235	0.410	0.070	
Detailed	マルチ	gemma3-12b-it	<u>0.345</u>	0.335	0.230	0.090	0.287
Detailed	マルチ	Swallow-9b	0.280	0.255	0.435	0.030	
Detailed	マルチ	Llama-3.1-8B	0.255	0.245	0.405	0.095	
JSON	シングル	gpt-oss-20b	0.370	0.255	0.350	0.025	
JSON	シングル	Qwen3-14B	0.345	0.340	0.275	0.040	
JSON	シングル	gemma3-12b-it	<u>0.405</u>	0.425	0.110	0.060	0.353
JSON	シングル	Swallow-9b	0.285	0.435	0.165	0.115	
JSON	シングル	Llama-3.1-8B	0.360	0.340	0.160	0.140	
DPO	マルチ	Qwen3-14B	<u>0.420</u>	0.375	0.140	0.065	-

as-a-Judge による自動評価を行った。GPT-5.1 は reasoning effort を medium とした。ユーザモデルは gemma3-12b を利用した。100 件の対話ログを、評価用データセットを使用してユーザモデルと対話システムを対話させることで収集した。合計 200 文対が評価対象になる。発話例は図 1 の左側に示し、全文を図 4 に示す。LLM-as-a-Judge では gpt-oss-20b を reasoning effort: medium, temperature: 0.3 で利用し、3 回評価を生成して多数決によりラベルを決定した。

人手評価 自動評価の妥当性を検証するために、自動評価で収集した 200 件の文対のうち、ランダムに選択した 100 件を人手評価した。Qwen3-14B を対象とした Baseline, Detailed, JSON, DPO の 4 手法と、GPT-5.1 の Baseline の合計 5 例を対象とした。3 人の日本語母語話者の大学生に対して 500 件ずつアノテーションを依頼した。アノテーション指示は LLM-as-a-Judge のプロンプトに倣って作成した。

4.2 実験結果

自動評価 自動評価結果を表 1 に示す。手法ごとの平均 SameRate を確認すると、Baseline, Detailed, JSON の順にスコアが向上している。JSON は Detailed と指示が同等であり、シングルターンプロンプトにすることで、制御性が向上したと考えられる。Qwen3-14B においては DPO 手法が最高のスコアを達成している。一方で GPT-5.1 のスコアは低い。ChatHarderRate が高いため、モデルの高難易度

表 2 人手評価の結果

手法	モデル	human consensus	LLM-as-a-Judge
Baseline	GPT-5.1	0.20	0.12
Baseline	Qwen3-14B	0.56	0.30
Detailed	Qwen3-14B	0.57	0.24
JSON	Qwen3-14B	0.75	0.37
DPO	Qwen3-14B	0.80	0.44

の発話を選ぶ傾向が影響していると考えられる。

人手評価 アノテーションの結果を表 2 に示す。3 人のアノテータ間の Fleiss' Kappa は 0.36, 人手の多数決後のラベルと LLM-as-a-Judge の間の Cohen's Kappa は 0.31 であり、タスクの主観性を考えると妥当な値となった。また、スピアマン順位相関は 0.9 となり傾向が一致していることがわかった。Qwen3-14B の Baseline と Detailed の 2 つの結果についてアノテータと LLM-as-a-Judge の間で順位が反転しているが、他の順位は一貫している。

5 おわりに

本研究では、第二言語学習者のトピックごとの発話難易度変化に対して近接発達領域内の発話を提供して学習を支援することを目標に、ユーザモデルを利用した評価フレームワークを提案した。JSON 手法と選好ペアで学習した DPO 手法のそれぞれで性能の向上が確認できた。また、人手評価と比較することで LLM-as-a-Judge の妥当性を確認できた。今後はより詳細なエラー分析に取り組む。

謝辞

本研究は、JSPS 科研費（若手研究，課題番号：JP24K20840）の助成を受けて実施した。

参考文献

- [1] Mostafa Papi and Hassan Khajavy. Second Language Anxiety: Construct, Effects, and Sources. **Annual Review of Applied Linguistics**, Vol. 43, p. 127–139, 2023.
- [2] S.D. Krashen. **The Input Hypothesis: Issues and Implications**. Longman, 1985.
- [3] Meiqing Jin, Liam Dugan, and Chris Callison-Burch. Controlling Difficulty of Generated Text for AI-Assisted Language Learning. **arXiv:2506.04072**, 2025.
- [4] Mina Almasi and Ross Deans Kristensen-McLachlan. Alignment Drift in CEFR-Prompted LLMs for Interactive Spanish Tutoring. In **Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 70–88, 2025.
- [5] Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. Grammar Control in Dialogue Response Generation for Language Learning Chatbots. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics**, pp. 9820–9839, 2025.
- [6] Xuyan Qiu. Functions of Oral Monologic Tasks: Effects of Topic Familiarity on L2 Speaking Performance. **Language Teaching Research**, Vol. 24, No. 6, pp. 745–764, 2020.
- [7] Mahmoud Abdi Tabari and Yizhou Wang. Assessing Linguistic Complexity Features in L2 Writing: Understanding Effects of Topic Familiarity and Strategic Planning Within the Realm of Task Readiness. **Assessing Writing**, Vol. 52, p. 100605, 2022.
- [8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.
- [9] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In **Proceedings of the 29th Symposium on Operating Systems Principles**, p. 611–626, 2023.
- [10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 38–45, 2020.
- [11] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical Analysis of Training Strategies of Transformer-Based Japanese Chat-Chat Systems. In **2022 IEEE Spoken Language Technology Workshop**, pp. 685–691, 2023.
- [12] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. **arXiv:2508.10925**, 2025.
- [13] Qwen Team. Qwen3 Technical Report. **arXiv:2505.09388**, 2025.
- [14] Gemma Team. Gemma 3 Technical Report. **arXiv:2503.19786**, 2025.
- [15] 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.
- [16] Llama Team. The Llama 3 Herd of Models. **arXiv:2407.21783**, 2024.
- [17] Rina Miyata, Toru Urakawa, Hideaki Tamori, and Tomoyuki Kajiwaru. Unsupervised Sentence Readability Estimation Based on Parallel Corpora for Text Simplification. In **Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 499–504, 2025.
- [18] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. MATCHA: 専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. **自然言語処理**, Vol. 31, No. 2, pp. 590–609, 2024.
- [19] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 2526–2547, 2025.
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In **International Conference on Learning Representations**, 2022.

A 手法プロンプト

各手法のプロンプトを図 3 に示す。

<p>Baseline</p> <p>次のルールに従って会話してください：</p> <ol style="list-style-type: none">必ず日本語で話してください。ユーザーのトピックに沿って話してください。新しい話題を提起したり、勧めたりしないでください。「さよなら」や「また会いましょう」などの対話を終わらせるような発話をしないでください。出力は発話内容のみとし、説明文や記号（例：「：」「※」「---」「#」改行コードなど）は含めないでください。実際の会話を想定し、必ず短い日本語で話してください。1発話あたり1から3文で話してください。相手の発話に対して、会話の流れに合わせて自然に相づちや質問をしてください。ただし毎回質問する必要はなく、状況に応じて適切な反応をしてください。
<p>Detailed</p> <p>あなたは日本語学習者の会話相手です。相手の日本語能力に合わせて、自然な会話を続けてください。</p> <p>[基本ルール]</p> <p>次のルールに従ってください：</p> <ol style="list-style-type: none">必ず日本語で話してください。ユーザーのトピックに沿って話してください。新しい話題を提起したり、勧めたりしないでください。「さよなら」や「また会いましょう」などの対話を終わらせるような発話をしないでください。実際の会話を想定し、必ず短い日本語で話してください。1発話あたり1から3文で話してください。相手の発話に対して、会話の流れに合わせて自然に相づちや質問をしてください。ただし毎回質問する必要はなく、状況に応じて適切な反応をしてください。 <p>[難易度調整のガイドライン]</p> <p>次のガイドラインに従って、相手の日本語能力に合わせて会話の難易度を調整してください：</p> <ol style="list-style-type: none">相手の言い方をそのまま真似るオウム返しはしないでください。相手の直前の発話に注目し、その語彙レベルに合わせた語彙を使って話してください。相手が日常会話でよく出現する単語を使っている場合は、同様のレベルの単語を選んでください。より高度な抽象的な単語を使っている場合は、少し難易度を上げた語彙を使用してください。相手の直前の発話に注目し、その文法・文構造レベルに合わせた表現を使って話してください。相手が修飾構造や接続が複雑でない簡単な文法を使っている場合は、同様のレベルの文法を使用してください。より複雑な文法や長い文を使っている場合は、それに合わせて複雑な修飾構造や接続を使ってください。相手の発話とだいたい同じ長さで発話してください。
<p>JSON</p> <p>あなたは日本語学習者の会話相手です。相手の日本語能力に合わせて、自然な会話を続けてください。</p> <p>[対話履歴]</p> <p>あなたはCHATです。USERと対話しています。 USERの発話に対する応答を生成してください。</p> <p>{history}</p> <p><Detailedの[基本ルール]と[難易度調整のガイドライン]を挿入></p> <p>[出力フォーマット]</p> <p>直前の{last_user}</p> <p>直前のUSER発話に対する応答を、JSON形式で出力してください。</p> <p>次のJSONだけを必ず出力してください。説明や考察は一切書かないでください。</p> <pre>{ "utterance": <実際の発話内容> }</pre>

図 3 手法ごとのプロンプト例

B 出力例

Qwen3-14B の Baseline 手法における出力例を図 4 に示す。

<p>CHAT(ダミー): こんにちは。最近どうですか？</p> <p>USER: こんにちは。ええ、元気です。あなたは？ 私は犬の散歩に行きました。</p> <p>CHAT: こんにちは。元気そうでよかったです。犬の散歩、楽しかったですか？</p> <p>USER: はい、とても楽しかったです。公園で、他の犬と遊びました。あなたは、何かしましたか？</p> <p>CHAT: いいですね、他の犬と遊ぶのは犬も喜ぶと思います。私は昨日、本を読んだり、ちょっとだけテレビを見たりしました。</p> <p>USER: ええ、そうですか。本は何を読みましたか？ 私は、音楽が好きです。クラシック音楽を聞くのが好きです。</p> <p>CHAT: 最近小説を読んでいることが多いです。クラシック音楽は落ち着くんですね。どんな曲が好きですか？</p> <p>USER: そうですね、落ち着きますよね。私は、子どもとバスケットボールをすることに夢中です。公園で一緒にシュート練習をする時、何かごほうびはどうしていますか？</p> <p>CHAT: 子どもが頑張ったときには、お菓子やお気に入りのスナックをあげていますよ。子ども喜んで練習に励みますね。</p> <p>USER: それはいいですね！ どのお菓子が好きですか？ 私は、子どもに何かごほうびをあげるとき、ポイント制にしようと思っています。</p> <p>CHAT: ポイント制はいいですね。子ども目標に向かって頑張る気がするかもしれませんよ。どんなポイントにしようか計画していますか？</p> <p>USER: そうですね、目標があると頑張れますよね。例えば、シュートが決まったらポイント、ドリブルを上手にできたらポイント、という風に考えています。どう思いますか？</p> <p>CHAT: その方法なら、子ども楽しみながら練習できそうですね。ポイントをためた後のお楽しみも用意しておけば、もっとやる気が出るかもしれませんよ。</p>	<p>前半 (ターン1~3)</p> <p>話題: クラシック音楽のコンサート 難易度: N 5</p> <p>human consensus: same LLM as a Judge: same</p> <p>後半 (ターン4~6)</p> <p>話題: 子どもと楽しむバスケットボール 難易度: N 3</p> <p>human consensus: same LLM as a Judge: same</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

図 4 出力例