A Dialogue System for Second Language Learning with Dynamic Adaptation to In-Dialogue Difficulty Changes

Taku Morioka [†] **Junya Takayama** [†] **Tomoyuki Kajiwara** [†‡]
[†] Ehime University [‡] The University of Osaka
{morioka@ai., takayama@ai., kajiwara@} cs.ehime-u.ac.jp

Abstract

We propose a dialogue system for second language learning that dynamically adjusts the difficulty of its utterances by adapting to changes in the learner's utterance difficulty across dialogue topics. Recent studies have advanced second language learning dialogue systems based on large language models (LLMs) grounded in the Input Hypothesis. However, although learners' language proficiency varies based on their familiarity with a given topic, conventional fixed-difficulty dialogue systems do not adequately account for this variability. In this study, we propose a dialogue system for second language learning that minimizes the difficulty gap between the system and a user model. The user model dynamically varies utterance difficulty based on the topic during dialogue. Our LLM-based dialogue system controls utterance difficulty by prompt engineering and Direct Preference Optimization (DPO). Experimental results in both Japanese and English demonstrate that the proposed method is effective in tracking utterance difficulty for second language learners.

1 Introduction

The increasing globalization of communication has led to a growing demand for second language learning. However, learners often have limited access to opportunities for conversational practice. This is due not only to the high financial cost but also to learners' anxiety when speaking a second language with others (Papi and Khajavy, 2023). Recently, large language models (LLMs) have become widely available and provide conversational interfaces, which may reduce the cost of conversational practice and alleviate learners' anxiety in interpersonal communication.

In second language learning, it is crucial that the input provided to learners remains within a comprehensible range that is neither too easy nor too difficult (Krashen, 1985). In prior work on dialogue systems for second language learners, methods have been proposed to control system outputs in order to satisfy a predetermined, fixed level of utterance difficulty (Jin et al., 2025; Almasi and Kristensen-McLachlan, 2025; Glandorf et al., 2025). However, it has been shown that learners' utterance difficulty varies depending on their familiarity with the topic (Qiu, 2020; Abdi Tabari and Wang, 2022), and existing dialogue systems based on fixed-difficulty control (Jin et al., 2025; Almasi and Kristensen-McLachlan, 2025; Glandorf et al., 2025) do not account for this variation. To provide comprehensible input that is appropriate for learning, it is necessary to dynamically adapt to the learner's utterance difficulty during the dialogue. In this setting, the learner's current proficiency relative to the topic may be approximated from the difficulty of their utterances.

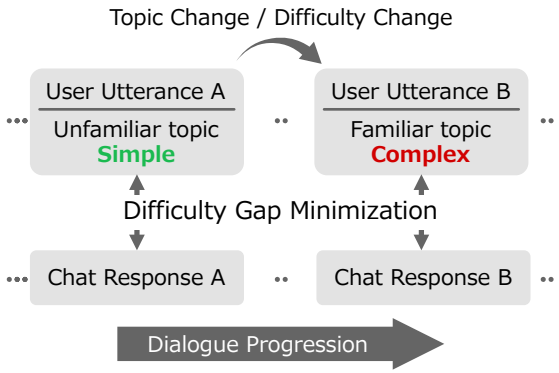

Figure 1: Overview of a dialogue system that adapts to changes in user utterance difficulty. The order of simple and complex user utterances is randomly determined.

In this study, we aim to develop a dialogue system that adapts to the learner's varying utterance difficulty and provides appropriate input (Figure 1). We also propose an evaluation task and methods for controlling the difficulty of system responses. In the evaluation task, we generate dialogue histories using a user model constructed with an LLM

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 170–187
August 2–5, 2026. ©2026 SIGDIAL

and assess the difficulty difference for each pair of utterances. The user model randomly switches between topics with high and low familiarity in the first and second halves of the dialogue, simulating changes in utterance difficulty. We extract utterance pairs from the generated dialogue histories and evaluate the system's ability to adapt to utterance difficulty from both lexical and syntactic perspectives. These aspects are considered key components of linguistic complexity in second language learning research (Bulté and Housen, 2012). Our evaluation task is related to the self-chat approach, which assesses dialogues between LLMs (Almasi and Kristensen-McLachlan, 2025), but it is novel in that the user model's utterance difficulty changes dynamically during the dialogue.

We propose methods for controlling utterance difficulty using prompts and a DPO-based approach (Rafailov et al., 2023) to improve a dialogue system's ability to adapt to varying utterance difficulty. As prompt-based methods for controlling utterance difficulty, we propose two approaches: the Detailed method and the Single-Block method. The Detailed method assigns instructions to the system role and the dialogue history to the user and assistant roles in the LLM chat template, whereas the Single-Block method provides both the instructions and the dialogue history in a single user role. In the DPO-based method, utterances from multiple models are automatically evaluated, and preference pairs are created for training. In experiments in both Japanese and English, we showed the effectiveness of the proposed methods for multiple LLMs using automatic evaluation. In both languages, the proposed methods achieved higher average scores than simple dialogue instructions. The results showed that, for prompt-based control, the Single-Block method was effective, while DPO was effective when training was feasible. Furthermore, to verify the reliability of the automatic evaluation, we conducted a human evaluation in Japanese and observed a high Spearman rank correlation as well as a reasonable Cohen's κ.

The main contributions of this work are as follows:

1. We propose a novel evaluation task and response generation methods for dialogue systems that adapt to learners' varying utterance difficulty.

2. We propose prompt-based control methods (Detailed and Single-Block) and a DPO-based approach, and demonstrate their effectiveness through experiments in both Japanese and English.

3. We show the effectiveness of the proposed methods and the reliability of the evaluation using both automatic and human assessments.

Our code is available at `https://github.com/EhimeNLP/DiAL2`.

2 Related Work

With the advancement of large language models (LLMs), research on dialogue systems for second language learners has progressed rapidly. Many prior studies are grounded in the Input Hypothesis (Krashen, 1985), which posits that comprehensible input facilitates language acquisition, and aim to generate utterances tailored to learners' proficiency levels.

The Common European Framework of Reference for Languages (CEFR) is widely used as a benchmark for utterance difficulty. CEFR is an international standard that defines language proficiency across six levels (A1–C2), with higher levels corresponding to greater linguistic competence. In some cases, language-specific proficiency scales are employed. For example, in Japanese, the Japanese Language Proficiency Test (JLPT) serves as a commonly used benchmark. JLPT defines five levels (N5–N1), where N1 represents the highest proficiency. Prior work (Jin et al., 2025; Almasi and Kristensen-McLachlan, 2025; Glandorf et al., 2025) assumes a target learner proficiency level in advance and proposes dialogue systems that generate utterances corresponding to a designated difficulty level.

Almasi and Kristensen-McLachlan (2025) analyze adherence to CEFR-based prompting in Spanish and report a phenomenon termed alignment drift, in which compliance with CEFR instructions deteriorates as the number of dialogue turns increases. They evaluate difficulty control performance using readability metrics and measures of syntactic complexity. Some studies introduce Future Discriminators for Generation (FUDGE) (Yang and Klein, 2021) to explicitly control the generation process (Jin et al., 2025). FUDGE regulates decoding by employing a future discriminator that estimates the probability that the next token will lead to an output satisfying a desired target attribute. Jin et al. (2025) combine a lexical difficulty

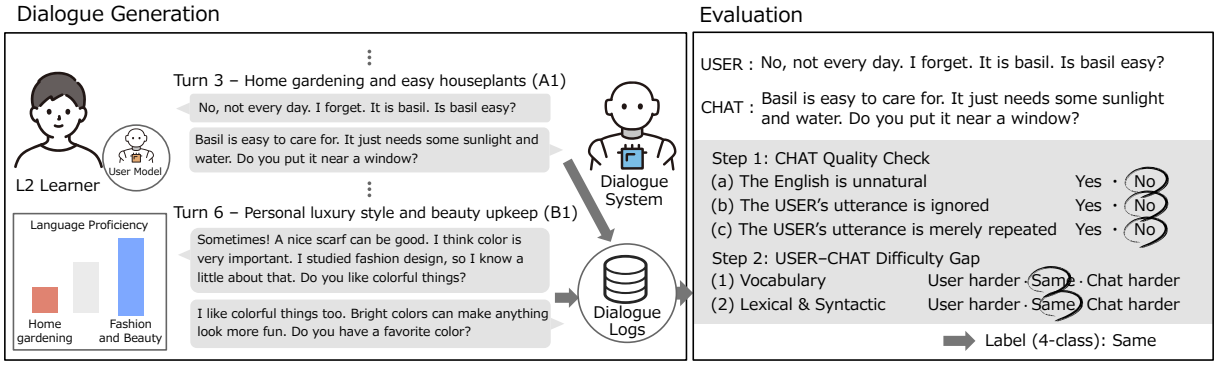

Figure 2: Overview of the evaluation task. Dialogues are collected from interactions between the user model and the dialogue system, and the difficulty gap is evaluated for the utterance pair at a target turn.

dictionary with BERT-based sentence difficulty estimation to evaluate the effectiveness of FUDGE for difficulty control in Japanese. Glandorf et al. (2025) train an automatic grammar detector and combine it with FUDGE to control the occurrence of specific grammatical constructions in generated outputs. They evaluate whether the instructed grammatical items appear in the generated outputs using a trained classifier. All of these studies assume a setting in which the target level is fixed in advance, and utterances are generated to conform to the given difficulty level. However, in real-world dialogue, a learner's utterance difficulty may vary depending on the topic and context. Consequently, fixed-difficulty control is insufficient to fully adapt to such fluctuations during interaction. In contrast to prior work, this study focuses on a dialogue system that dynamically adjusts its response difficulty according to the learner's utterances, and evaluates its ability to adapt to variations in the learner's utterance difficulty throughout the dialogue.

Self-chat, in which a user model and a system model engage in dialogue as LLMs, is commonly used to evaluate dialogue systems for second language learners (Almasi and Kristensen-McLachlan, 2025; Jin et al., 2025; Glandorf et al., 2025). Self-chat is a framework for evaluating a system model's control performance using dialogue histories generated through interactions between a user model and a system model. In this study, we extend the self-chat framework by introducing dynamic changes in the user model's utterance difficulty and use it to design our evaluation task.

3 Task Definition

We propose a novel evaluation task for measuring how well a dialogue system adapts to changes in the difficulty of a second language learner's utterances across different dialogue topics. The task is constructed in a self-chat setting using a user model that simulates a learner. An overview of the evaluation framework is shown in Figure 2. First, we select a large language model (LLM) that demonstrates strong adherence to difficulty-level instructions as the user model. We then construct a dataset consisting of dialogue topics, personas, and difficulty-level instructions provided to the user model. Each dialogue is conducted as a multi-turn interaction. By changing the dialogue topic and difficulty instructions given to the user model between the first and second halves, we create a scenario in which the user's utterance difficulty varies during the dialogue. The dialogue system interacts with the user model under this dynamic setting. Evaluation is performed at the turn level. From the final turn of both the first and second halves of the dialogue, we extract pairs of user-model utterances (USER) and system utterances (CHAT), and assign labels according to the evaluation criteria described below.

3.1 Evaluation Criteria

Step 1: Quality Assessment We first assess whether the CHAT utterance is suitable for difficulty comparison using three binary criteria. (a) unnatural Japanese, (b) ignoring the USER utterance, and (c) simple repetition of the USER utterance. If any criterion is positive, the pair is labeled *Reject* and excluded from Step 2.

Step 2: Difficulty Comparison We then compare USER and CHAT utterances in terms of lexical and syntactic complexity. For each dimension, we assign one of three labels: U (USER harder), S (same), or C (CHAT harder).

3.2 Label Aggregation

To facilitate analysis, we aggregate the five evaluation results into four final labels: *Reject*, *UserHarder*, *Same*, and *ChatHarder*. *Reject* is assigned if any Step 1 criterion is positive. *UserHarder* (resp. *ChatHarder*) is assigned when both perspectives are U (resp. C), or when one is U (resp. C) and the other is S. *Same* is assigned when both are S or when U and C labels conflict.

4 Proposed Method

We propose three difficulty control methods: two training-free prompt-based approaches (Detailed and Single-Block) and one training-based approach using DPO, together with a simple instruction-based baseline for comparison.

Baseline (comparison) Baseline uses only a simple dialogue instruction. Following the LLM chat template, the instruction prompt is placed in the system role, and the dialogue history is provided in alternating user and assistant roles. The Japanese prompt and its English translation are shown in Figures 16 and 17, respectively.

Prompt-Based Methods The Detailed method extends the baseline by explicitly instructing the system to match the user's vocabulary, syntax, and utterance length. The Japanese prompt and its English translation are shown in Figures 18 and 19, respectively. The Single-Block method provides similar instructions but places both the prompt and the dialogue history within a single user message. In addition, it enforces JSON-formatted output to avoid generating extraneous text. The Japanese prompt and its English translation are shown in Figures 20 and 21, respectively.

DPO We define responses with high difficulty alignment as preferred responses and optimize the LLM using Direct Preference Optimization (DPO) (Rafailov et al., 2023). Preference pairs are automatically constructed. For each dialogue history x, multiple response candidates are generated and evaluated using the automatic evaluation criteria described in Section 3.1. A penalty score is computed as the sum of quality violations and the number of difficulty labels that are not *Same*. When a candidate with zero penalty exists, it is selected as the positive example y^+, and the candidate with the highest penalty is selected as the negative example y^-. The model is then optimized using (x, y^+, y^-).

Model	Mean Error ↓	
	Ja	En
gpt-oss-20b	1.192	1.317
Qwen3-14B	1.386	1.565
gemma3-12b	**1.168**	**0.885**
Llama-3.1-8B	1.496	1.352
Swallow-9b	1.552	–

Table 1: Results of the user model selection experiment. Mean error represents the absolute difference between the instructed difficulty level and the estimated difficulty label of the generated utterance.

5 Evaluation Experiments

5.1 Experimental Setup

We evaluate the proposed methods in both Japanese and English using the evaluation task. We use vLLM (Kwon et al., 2023) for LLM inference.

User Model Selection To select the user model for the evaluation experiments, we evaluated the difficulty control ability of several LLMs. Difficulty was represented using the five-level JLPT scale for Japanese and the six-level CEFR scale for English. We sampled 200 dialogues with associated personas from JPersonaChat for Japanese (Sugiyama et al., 2023) and PersonaChat (Zhang et al., 2018) for English, using the first five utterances of each dialogue as context. For each dialogue history, we generated responses under different difficulty levels, resulting in 1,000 utterances per model for Japanese (200 × 5 levels) and 1,200 for English (200 × 6 levels). Following Miyata et al. (2025), we estimated utterance difficulty via pairwise classification. For pairwise comparison, we used ModernBERT (Warner et al., 2025) fine-tuned on MATCHA (Miyata et al., 2024) for Japanese and on Newsela-Auto (Jiang et al., 2020) for English. Further details are provided in Appendix A. We compared the following models: gpt-oss-20b[1] (OpenAI, 2025), Qwen3-14B[2] (Qwen Team, 2025), gemma3-12b[3] (Gemma Team, 2025), and Llama-3.1-8B[4] (Llama Team, 2024). For Japanese, we additionally included Swallow-9B[5] (Fujii et al., 2024) in the compari-

[1] https://huggingface.co/openai/gpt-oss-20b
[2] https://huggingface.co/Qwen/Qwen3-14B
[3] https://huggingface.co/google/gemma-3-12b-it
[4] https://huggingface.co/meta-llama/Llama-3.1-8B
[5] https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1

son. All models are publicly available LLMs with no more than 20B parameters. Mean Error was defined as the average absolute difference between the instructed and estimated difficulty labels and used to evaluate difficulty controllability. As shown in Table 1, gemma3-12b achieved the lowest mean error in both Japanese and English, and was therefore selected as the user model.

Construction of the User Model Instruction Dataset We construct a user instruction dataset to control the user model with respect to topic interest and utterance difficulty, consisting of (i) persona, (ii) dialogue themes, (iii) specific topics, and (iv) target difficulty labels. Personas were extracted from JPersonaChat (Sugiyama et al., 2023) for Japanese and from PersonaChat (Zhang et al., 2018) for English. Using each persona as input, we employed an LLM to automatically generate dialogue themes and topics likely or unlikely to interest the persona. Dialogue themes were generated using the GPT-5 API[6]. The generation prompts are shown in Figure 22 for Japanese and Figure 23 for the translated English version. Each dialogue is divided into two parts, where themes that the persona is likely or unlikely to be interested in are randomly assigned. This allows us to evaluate difficulty control under changes in topic interest within a single dialogue. As our target application is a dialogue system for second language learners, we use N3 and N5 for Japanese and B1 and A1 for English to avoid excessively large difficulty gaps while maintaining clear differences in proficiency levels. N3 was assigned to topics likely to interest the persona, and N5 to topics unlikely to interest them. In total, we constructed 4,897 training dialogues and 100 evaluation dialogues for Japanese, and 5,100 training dialogues and 100 evaluation dialogues for English.

Dialogue Generation To evaluate difficulty control, we generated six-turn dialogues between the user model and the target system. Different topics and difficulty levels were assigned to turns 1–3 and turns 4–6 to induce changes in the user model's behavior. To satisfy the user model's input format, dialogues were initiated with a dummy utterance: "こんにちは。最近どうですか？" in Japanese and "Hello! How are you these days?" in English. At turn 4, the topic and difficulty were explicitly switched, and to prevent influence from the previous context, the user model was not given dialogue history from the first three turns. User model prompts are provided in Appendix F.

Implementation of Methods For Baseline, Detailed, and Single-Block, we provided the corresponding prompts to the LLM. Since Swallow-9b does not support system prompts, the instruction prompt and the first turn utterance were included in the user prompt for the Baseline and Detailed methods. Because DPO training is computationally expensive, it was applied only to Qwen3-14B. First, initial dialogues were generated with Qwen3-14B using the Single-Block method, and additional response candidates were generated with the Baseline method using gpt-oss-20b, Llama-3.1-8B, and Swallow-9B (Japanese only). The candidate responses were automatically evaluated to construct 6,066 preference pairs for Japanese and 6,698 for English. The model was trained using LoRA (Hu et al., 2022) via the trl library[7]. The LoRA hyperparameters are provided in Appendix D. During inference, the Baseline prompt was used.

Automatic Evaluation We conducted automatic evaluation using LLM-as-a-Judge with the following models: gpt-oss-20b, Qwen3-14B, gemma3-12b, Swallow-9B (Japanese only), Llama-3.1-8B, and GPT-5.1[8](reasoning effort: medium). We used gemma3-12b as the user model and collected 100 dialogues using the evaluation instruction dataset. One utterance pair was extracted from both the first and second halves of each dialogue, resulting in 200 pairs. For automatic evaluation, we used the largest local model, gpt-oss-20b, with reasoning effort set to medium and temperature set to 0.3. Each utterance pair was evaluated three times, and the final label was determined by majority voting. For the ensemble, the majority label among the four classes was selected; if all labels differed, the Same label was assigned. For Japanese, the prompts shown in Figures 10, 11, and 12 were used for Step 1, lexical evaluation, and syntactic evaluation, respectively. The English prompts were translated from the Japanese prompts. However, because translation may alter the intent of the few-shot examples, we verified the effect of removing them in Japanese (Appendix B). The prompts used are shown in Figures 13, 14, and 15.

[6] https://platform.openai.com/docs/models/gpt-5, accessed between 2025/12/07 and 2026/01/30

[7] https://github.com/huggingface/trl

[8] https://platform.openai.com/docs/models/gpt-5.1, accessed between 2025/12/07 and 2026/02/28

Method	Model	Japanese			English		
		Same↑	User>Chat↓	Chat>User↓	**Same**↑	User>Chat↓	Chat>User↓
Baseline	GPT-5.1	0.100	0.065	0.750	0.185	0.000	0.760
	gpt-oss	0.270	0.155	0.550	0.560	0.010	0.420
	Qwen3	0.205	0.265	0.505	0.670	0.030	0.260
	gemma3	0.375	0.380	0.155	0.680	0.050	0.225
	Llama3.1	0.325	0.260	0.310	0.155	0.010	0.795
	Swallow	0.370	0.335	0.230	–	–	–
Detailed	GPT-5.1	0.235	0.130	0.590	0.455	0.005	0.490
	gpt-oss	0.285	0.260	0.435	0.625	0.020	0.335
	Qwen3	0.280	0.270	0.405	0.740	0.015	0.230
	gemma3	0.400	0.365	0.175	0.695	0.020	0.235
	Llama3.1	0.225	0.220	0.435	0.325	0.020	0.610
	Swallow	0.270	0.240	0.475	–	–	–
S-Block	GPT-5.1	0.235	0.150	0.555	0.340	0.005	0.600
	gpt-oss	0.315	0.300	0.355	0.710	0.055	0.220
	Qwen3	0.320	0.380	0.260	0.765	0.030	0.195
	gemma3	0.360	0.390	0.150	0.715	0.045	0.205
	Llama3.1	0.285	0.335	0.195	0.710	0.040	0.165
	Swallow	0.320	0.475	0.130	–	–	–
DPO	Qwen3	**0.420**	0.295	0.240	**0.795**	0.080	0.090

Table 2: Automatic evaluation results on 200 utterance pairs for Japanese and English. S-Block denotes the Single-Block method. User>Chat and Chat>User indicate the UserHarder and ChatHarder labels, respectively.

Method	Mean Same Rate ↑	
	Ja	En
Baseline	0.274	0.450
Detailed	0.283	0.568
Single-Block	**0.306**	**0.648**

Table 3: Average Same rates for Baseline, Detailed, and Single-Block in the automatic evaluation.

Human Evaluation on Japanese To validate the automatic evaluation, we randomly sampled 100 utterance pairs from the 200 pairs collected via automatic evaluation and conducted a human evaluation. The evaluation covered five conditions: Baseline, Detailed, Single-Block, and DPO for Qwen3-14B, and Baseline for GPT-5.1. Three native Japanese-speaking university students each annotated 500 utterance pairs. The final label was determined by majority voting; if all annotators assigned different labels, *Same* was assigned. The annotation guidelines were designed to be equivalent to the LLM-as-a-Judge prompt.

5.2 Experimental Results

5.2.1 Automatic Evaluation

Automatic evaluation results are shown in Table 2. Higher Same rates are desirable, whereas lower values are preferred for the other metrics. Except for Detailed and Single-Block in English, gemma3-12b-it achieves the best performance for most settings, likely because its self-chat partner is also gemma3-12b-it. Overall, DPO with Qwen3-14B achieves the highest Same rate in both Japanese and English. Label conflicts do not account for this gain (Appendix E). For Qwen3-14B, DPO increased the Same rate compared to the Baseline, improving from 0.205 to 0.420 in Japanese and from 0.670 to 0.795 in English. GPT-5.1 shows relatively low Same rates in both languages, mainly due to the high proportion of ChatHarder labels.

The average Same rates for each method are shown in Table 3. Performance improves in the order of Baseline, Detailed, and Single-Block, suggesting that the single-turn prompting format in Single-Block improves controllability. These results indicate that DPO is effective when training

Method	Model	Human Evaluation	LLM-as-a-Judge
Baseline	GPT-5.1	0.35	0.13
Baseline	Qwen3-14B	0.80	0.24
Detailed	Qwen3-14B	0.68	0.31
Single-Block	Qwen3-14B	0.77	0.32
DPO	Qwen3-14B	**0.85**	**0.44**

Table 4: Human evaluation results on 100 utterance pairs.

Method	mean			over		
	user	chat	diff ↓	user	chat	diff ↓
Baseline	1.648	1.790	0.142	0.183	0.239	0.057
Detailed	1.608	1.588	-0.020	0.178	0.170	-0.009
Single-Block	1.626	1.616	**-0.010**	0.185	0.188	**0.002**
DPO	1.626	1.694	0.067	0.177	0.201	0.024

Table 5: Lexical-level analysis for each method of Qwen3-14B. mean denotes the average difficulty, and over denotes the proportion of words exceeding the target difficulty.

is feasible, whereas Single-Block is suitable when only prompt-based control is used. The effectiveness of DPO was also confirmed in experiments where the target difficulty levels instructed to the user model in Japanese were changed to the N4–N2 and N3–N1 pairs (Appendix C).

5.2.2 Human Evaluation

Human evaluation results are shown in Table 4. Although the Baseline score for Qwen3-14B is relatively high in human evaluation, the overall trends are similar to those observed in the automatic evaluation. At the instance level, Inter-annotator agreement was moderate: Fleiss' κ among the three annotators was 0.31 and the average pairwise Cohen's κ was 0.33. Cohen's κ between human labels and LLM-as-a-Judge was 0.24. This suggests that exact label matching for individual utterance pairs is only moderate, which is reasonable given the subjective nature of difficulty judgments. By contrast, at the level of overall model comparison, the agreement was stronger: the Spearman rank correlation was 0.70, indicating similar ranking tendencies. These results provide support for the validity of the automatic evaluation.

6 Analysis

6.1 Statistical Analysis

We analyze Japanese dialogues generated by Qwen3-14B to examine the behavior of the proposed methods. Among the 200 dialogues, 100 use an N5-to-N3 difficulty transition, while the remaining 100 use an N3-to-N5 transition. All six turns are analyzed from the perspectives of vocabulary and syntactic structure.

Method	user	chat	diff ↓
Baseline	1.495	1.710	0.215
Detailed	1.492	1.621	0.129
Single-Block	1.479	1.649	0.171
DPO	1.465	1.581	**0.116**

Table 6: Syntactic complexity (MHD) analysis for each method of Qwen3-14B.

Vocabulary For vocabulary analysis, we use jlpt-anki-decks[9], a dictionary that assigns JLPT labels to words. The extracted vocabulary consists of 2,683 N1 words, 1,732 N2 words, 2,101 N3 words, 660 N4 words, and 702 N5 words. Each sentence was tokenized using SudachiPy (Takaoka et al., 2018) in mode C, and tokens found in the JLPT dictionary were counted according to their difficulty level. The average difficulty was computed as $(N5\times1+N4\times2+N3\times3+N2\times4+N1\times5)/total$, where $total$ is the sum of counts from N5 to N1. We also computed the Over Rate, defined as the proportion of words exceeding the target difficulty (N4–N1 when the target is N5, and N2–N1 when the target is N3). These values are computed for

[9] https://github.com/chyyran/jlpt-anki-decks

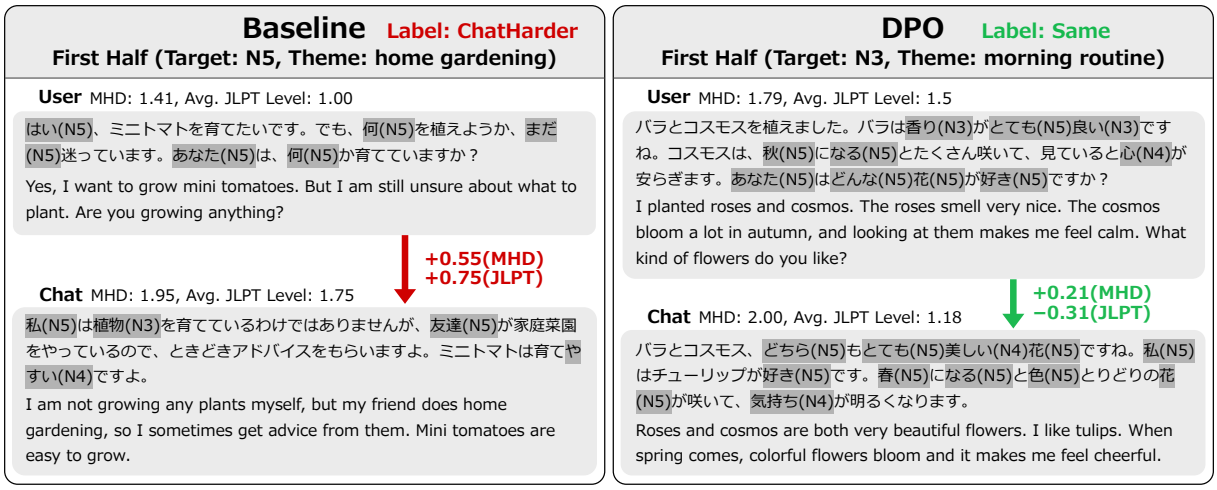

Figure 3: Example outputs from Qwen3-14B with the Baseline and DPO methods. The Baseline example is labeled ChatHarder, and the DPO example is labeled Same by both LLM-as-a-Judge and human evaluation.

both user and system utterances, and their differences are shown in Table 5. Baseline yields the worst scores for both the average difficulty difference (mean) and the proportion of words exceeding the target difficulty (over), whereas the proposed methods achieve better scores. Moreover, the Single-Block method achieves the best scores for both mean and over. Compared with Baseline, the Detailed, Single-Block, and DPO methods demonstrate improved performance. Baseline shows the largest difficulty gap between user and system utterances for both mean and over, whereas the proposed methods reduce this gap. Among them, Single-Block achieves the smallest differences, indicating the best lexical alignment with the user utterances.

Syntactic Structure To evaluate syntactic complexity, we compute the Mean Hierarchical Distance (MHD). MHD (Jing and Liu, 2015) measures the average dependency distance from the main predicate. We adopt this metric as it has been shown to be effective for measuring the complexity of Japanese sentences (Komori et al., 2019). Similar to the lexical analysis, we compute the MHD of the user and system utterances and their difference, and summarize the results in Table 6. Compared with Baseline, the Detailed, Single-Block, and DPO methods yield smaller differences, indicating improved syntactic alignment with the user utterances. DPO achieves the smallest difference.

6.2 Case Study

From the outputs of Qwen3-14B, we extracted two representative examples: one labeled ChatHarder under the Baseline method and one labeled Same under the DPO method, both agreed upon by human evaluation and LLM-as-a-Judge. These examples are shown in Figure 3. Under the Baseline method, the Chat utterance tends to be longer and more complex than the User utterance. The difference in MHD scores is large (0.55). In contrast, the absolute differences in MHD and JLPT vocabulary difficulty between User and Chat utterances are smaller under the DPO method, indicating improved alignment compared with Baseline.

7 Conclusion

We proposed an evaluation task and difficulty control methods for a second language learning dialogue system that adapts to dynamically changing learner utterance difficulty during dialogue. Experiments in both Japanese and English show that the proposed method outperforms a simple instruction-based baseline on average. Among the prompt-based approaches, Single-Block provides the most stable performance, while DPO achieves the best overall results when training is feasible. Human evaluation shows a ranking trend similar to the automatic evaluation, providing support for the proposed evaluation framework. These results indicate the potential of dialogue systems that follow changes in learner utterance difficulty depending on the topic and provide appropriately leveled responses. Future work includes evaluation with real learners, extension to more diverse dialogue settings, and improvement of automatic difficulty evaluation methods.

Limitations

Our study has several limitations. First, our evaluation relies on a simulated user model in a self-chat setting. While this framework enables low-cost and scalable evaluation, it may not fully capture the diversity and behaviors of real language learners. Therefore, future work should evaluate the proposed method with real language learners.

Second, due to the high training cost of DPO, we apply this method only to Qwen3-14B.

Third, we validate the evaluation task through human evaluation only in Japanese and do not directly examine its relationship with human evaluation in English. To avoid the influence of translation errors in the Japanese prompts, we analyze the effect of removing few-shot examples when introducing automatic evaluation in English (Appendix B). The English evaluation prompt is constructed by translating a Japanese prompt without few-shot examples, thereby minimizing cross-lingual effects.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24K20840.

References

Mahmoud Abdi Tabari and Yizhou Wang. 2022. Assessing Linguistic Complexity Features in L2 Writing: Understanding Effects of Topic Familiarity and Strategic Planning Within the Realm of Task Readiness. *Assessing Writing*, 52:100605.

Mina Almasi and Ross Deans Kristensen-McLachlan. 2025. 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*, pages 70–88.

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *First Conference on Language Modeling*.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv:2503.19786*.

Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. 2025. 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*, pages 9820–9839.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the Association for Computational Linguistics*.

Meiqing Jin, Liam Dugan, and Chris Callison-Burch. 2025. Controlling Difficulty of Generated Text for AI-Assisted Language Learning. *arXiv:2506.04072*.

Yingqi Jing and Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 161–170.

Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2019. Examining MDD and MHD as Syntactic Complexity Measures with Intermediate Japanese Learner Corpus Data. In *Proceedings of the Fifth International Conference on Dependency Linguistics*, pages 130–135.

Stephen D. Krashen. 1985. *The Input Hypothesis: Issues and Implications*. Longman.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Rina Miyata, Hyuga Koretak, Hiroki Yamauchi, Daiki Yanamoto, Tomoyuki Kajiwara, Takashi Ninomiya, and Yasuhiro Nishiwaki. 2024. MATCHA: Parallel Corpus for Japanese Text Simplification Based on Professionally Simplified Articles. *Journal of Natural Language Processing*, 31(2):590–609.

Rina Miyata, Toru Urakawa, Hideaki Tamori, and Tomoyuki Kajiwara. 2025. 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*, pages 499–504.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv:2508.10925*.

Mostafa Papi and Hassan Khajavy. 2023. Second Language Anxiety: Construct, Effects, and Sources. *Annual Review of Applied Linguistics*, 43:127–139.

Xuyan Qiu. 2020. Functions of Oral Monologic Tasks: Effects of Topic Familiarity on L2 Speaking Performance. *Language Teaching Research*, 24(6):745–764.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv:2505.09388*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical Analysis of Training Strategies of Transformer-Based Japanese Chit-Chat Systems. In *2022 IEEE Spoken Language Technology Workshop*, pages 685–691.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: A Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

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. 2025. 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*, pages 2526–2547.

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, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 38–45.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation with Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

A Classifier for User Model Selection

We describe the training of a classifier used for user model selection, which takes two sentences as input and predicts which one is more difficult. For both Japanese and English, the classifier was based on ModernBERT (Warner et al., 2025). For Japanese, we used MATCHA (Miyata et al., 2024). As preprocessing, we removed English expressions in parentheses and converted full-width alphabet characters to half-width. The data were randomly split into 14,000 training examples, 1,000 development examples, and 1,000 test examples. For English, we used the official Newsela-Auto split (Jiang et al., 2020): 394,300 training, 43,317 development, and 44,067 test examples. The classifier was trained for five epochs for both Japanese and English. The test accuracy was 0.93 for Japanese and 0.96 for English. We use HuggingFace Transformers (Wolf et al., 2020) for training and inference.

B Effect of Removing Few-Shot Examples from the Evaluation Prompt in Japanese

When translating the Japanese evaluation prompt into English, the intended meaning of the few-shot examples may not be fully preserved. Therefore, we translated the Japanese prompt without the few-shot section into English and used it in the experiments. In this section, we investigate the effect of removing the few-shot section from the Japanese evaluation prompt. The experiment was conducted under the same settings as the human evaluation (Section 5.1). Table 7 shows the results. Although removing the few-shot examples leads to a decrease in Cohen's κ, the impact is minor. In contrast, Spearman's ρ remains unchanged. Therefore, removing the few-shot prompt results in only a slight decrease in correlation, suggesting that translating and using it in English does not pose a significant problem.

C Experiment with Changes in Target Utterance Difficulty

In the Japanese experiments (Section 5), we used the N5–N3 difficulty pair. In this section, we conduct additional experiments using other difficulty pairs, namely N4–N2 and N3–N1. The only difference from Section 5 is the difficulty specification in the user model instruction dataset. We conducted experiments with four methods using Qwen3-14B. The results are shown in Table 8. DPO achieves

Method	Model	Human Evaluation	LLM-as-a-Judge	LLM-as-a-Judge (w/o few-shot)
Baseline	GPT-5.1	0.35	0.13	0.05
Baseline	Qwen3-14B	0.80	0.24	0.18
Detailed	Qwen3-14B	0.68	0.31	0.28
S-Block	Qwen3-14B	0.77	0.32	0.33
DPO	Qwen3-14B	**0.85**	**0.44**	**0.37**
Spearman's ρ		–	0.70	0.70
Cohen's κ		–	0.24	0.21

Table 7: Human evaluation results for 100 utterance pairs and the results of LLM-as-a-Judge with and without few-shot examples. Spearman's ρ and Cohen's κ are computed between the human evaluation and each LLM-as-a-Judge setting.

Method	N5–N3 ↑	N4–N2 ↑	N3–N1 ↑
Baseline	0.205	0.200	0.220
Detailed	0.280	0.275	0.180
S-Block	0.320	0.210	0.170
DPO	**0.420**	**0.340**	**0.230**
avg	0.306	0.256	0.200

Table 8: Proportion of the *Same* label under different target difficulties of the user model. Experiments were conducted with four methods of Qwen3-14B in Japanese.

Hyperparameter	Value
LoRA rank	8
LoRA alpha	32
LoRA dropout	0.05
Batch size	2
Epochs	3
Learning rate	5×10^{-5}
Beta	0.1
RPO alpha	1.0
Warmup ratio	0.03
Weight decay	0.01
Evaluation steps	500

Table 9: Hyperparameters used for DPO training.

the best performance across all difficulty pairs, indicating that the model can generalize to adapt to the partner's difficulty even when the preference dataset is constructed using the N5–N3 pair. The Baseline also maintains its performance even when the target difficulty increases. This may be because LLMs are pretrained on native-level text, allowing them to achieve reasonable performance even without additional control. Furthermore, as the target difficulty increases, the average score decreases, suggesting that adapting to the partner's utterance difficulty becomes more difficult at higher target difficulty levels.

D DPO Training Hyperparameters

Table 9 lists the hyperparameters used for training.

E Label Conflicts

To examine whether the high Same rate of DPO is an artifact of label conflicts, we analyze how often such conflicts occur. Label conflicts fall into two types: cross-dimension conflicts and three-way ties. A cross-dimension conflict occurs when, within a single run or annotation, the lexical and syntactic-complexity dimensions disagree—one labeled U and the other C—and the *Same* label is consequently assigned. A three-way tie occurs when, in majority voting across the three runs or three annotators, the votes split among the U, S, and C labels, and the *Same* label is assigned. Table 10 shows that the conflict rate is higher for Japanese than for English. Among all methods, DPO yields the lowest three-way-tie rate for Qwen3 in Japanese. Moreover, the conflict rate of DPO is at or below the overall average in every column except the English *merge* column. These results indicate that the effectiveness of DPO observed in our experiments is unlikely to be an artifact of label conflicts. The same trend holds for the human-evaluation conflict rates in Table 11: DPO exhibits the fewest three-way ties, so the impact of label conflicts on its observed effectiveness is negligible.

F User Model Prompts

For the user model instruction prompts, we used the prompts shown in Figure 4 for Japanese and

Method	Model	Japanese				English			
		Run 1	Run 2	Run 3	merge	Run 1	Run 2	Run 3	merge
Baseline	GPT-5.1	0.035	0.050	0.025	0.045	0.000	0.000	0.000	0.005
	gpt-oss	0.105	0.085	0.070	0.035	0.000	0.000	0.000	0.000
	Qwen3	0.030	0.055	0.035	0.025	0.000	0.005	0.000	0.005
	gemma3	0.075	0.070	0.070	0.040	0.000	0.000	0.000	0.000
	Llama3.1	0.050	0.040	0.015	0.055	0.000	0.000	0.000	0.005
	Swallow	0.035	0.035	0.050	0.055	–	–	–	–
Detailed	GPT-5.1	0.055	0.015	0.060	0.015	0.000	0.000	0.000	0.000
	gpt-oss	0.035	0.055	0.020	0.045	0.000	0.000	0.005	0.010
	Qwen3	0.030	0.020	0.030	0.045	0.000	0.000	0.000	0.000
	gemma3	0.025	0.050	0.005	0.035	0.000	0.000	0.000	0.000
	Llama3.1	0.030	0.030	0.045	0.015	0.000	0.000	0.000	0.000
	Swallow	0.050	0.045	0.055	0.035	–	–	–	–
S-Block	GPT-5.1	0.070	0.045	0.045	0.055	0.000	0.000	0.000	0.015
	gpt-oss	0.050	0.045	0.055	0.05	0.005	0.005	0.005	0.010
	Qwen3	0.030	0.020	0.010	0.035	0.000	0.000	0.000	0.000
	gemma3	0.010	0.025	0.010	0.015	0.000	0.000	0.000	0.005
	Llama3.1	0.005	0.015	0.015	0.025	0.005	0.000	0.005	0.010
	Swallow	0.025	0.015	0.025	0.010	–	–	–	–
DPO	Qwen3	0.030	0.040	0.030	0.010	0.000	0.000	0.000	0.005
Average		0.041	0.040	0.035	0.034	0.001	0.001	0.001	0.004

Table 10: Label conflict rates over 200 utterance pairs in the automatic evaluation. Run 1–3 report the proportion of cross-dimension conflicts in each of the three runs, where the lexical and syntactic-complexity dimensions disagree with one labeled U and the other C. *merge* reports the proportion of three-way ties at the majority-voting stage across the three runs, where the votes split among the U, S, and C labels.

Figure 5 for English. However, at the fourth turn, we appended the prompts shown in Figure 6 for Japanese and Figure 7 for English to clearly switch the topic. In addition, for the third and sixth turns used for evaluation, the user model was instructed to produce utterances consisting of three parts: a backchannel, the user's opinion or impression, and a question to the dialogue system, to avoid excessive variation in conditions across dialogue systems. For the evaluation turns, we appended the prompts shown in Figure 8 for Japanese and Figure 9 for English to the end of the instruction prompts.

G Additional Prompts

In this section, we present the prompts used in this study. For the evaluation prompts, we used Figures 10 for Step 1 and Figures 11 and 12 for Step 2 in Japanese. In English, we used Figures 13 for Step 1 and Figures 14 and 15 for Step 2. For the Baseline method, we used the prompts shown in Figure 16 for Japanese and Figure 17 for English. For the Detailed method, we used the prompts shown in Figure 18 for Japanese and Figure 19 for English. For the Single-Block method, we used the prompts shown in Figure 20 for Japanese and Figure 21 for English. For generating the user model instruction data, we used the prompts shown in Figure 22 for Japanese and Figure 23 for English.

Method	Model	Cross dimension conflicts			Three-way ties
		Annot. 1	Annot. 2	Annot.3	
Baseline	GPT-5.1	0.00	0.00	0.03	0.03
Baseline	Qwen3-14B	0.00	0.01	0.01	0.08
Detailed	Qwen3-14B	0.00	0.00	0.00	0.04
Single-Block	Qwen3-14B	0.00	0.00	0.00	0.03
DPO	Qwen3-14B	0.00	0.00	0.01	0.02

Table 11: Label conflict rates over 100 utterance pairs in the human evaluation. *Annot. 1–3* report the proportion of cross-dimension conflicts for each of the three annotators, where the lexical and syntactic-complexity dimensions disagree with one labeled U and the other C. *Three-way ties* reports the proportion of three-way ties in the majority vote across the three annotators, where the votes split among the U, S, and C labels.

あなたは第二言語である日本語を学習するJLPTレベル {target_level} の学習者として振舞ってください。
これから会話パートナーと、日本語の練習のために会話します。

あなたのペルソナは以下の通りです。
{persona_bulleted}

次のルールに従ってください：
1. これまでの対話の難易度は無視して、JLPTの {target_level} にふさわしい語彙と文法だけを使って話してください。
難しすぎる表現や上のレベルの表現は使わないでください。
ただし、子ども向けのとてもやさしい日本語にはしないでください。{target_level} の教科書に出てくるような文を目安にしてください。
2. 対話のテーマは「{current_theme}」です。以下のトピックを参考に会話してください。
{topics_bulleted}
注意：トピックの語彙や文法にとらわれず、必ずJLPT {target_level} にふさわしい表現を使ってください。
3. 必ず日本語で話してください。
4. 必ず会話を続けてください。ターン数が経過したからと会話を終了しないでください。
5. 「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。
6. 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。
7. 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。
8. 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。

Figure 4: Main user instruction prompt in Japanese

Please act as an English learner at CEFR {target_level}, where English is your second language.
From now on, you will have a conversation with a partner to practice English.

Your persona is as follows.
{persona_bulleted}

Please follow the rules below.
1. Ignore the difficulty of the conversation so far, and speak using only vocabulary and grammar appropriate for CEFR {target_level}.
Do not use expressions that are too difficult or above this level.
However, do not use very easy, childlike English. Use sentences like those found in a textbook for {target_level}.
2. The theme of the conversation is “{current_theme}”. Please talk with reference to the following topics.
{topics_bulleted}
Note: Do not stick to the vocabulary or grammar of the topics. Always use expressions appropriate for CEFR {target_level}.
3. Always speak in English.
4. Always continue the conversation. Do not end it just because many turns have passed.
5. Do not use expressions that end the conversation, such as “goodbye” or “see you”.
6. Output only the spoken content. Do not include explanations or symbols (for example, “:”, “※”, “---”, “#”, line breaks, etc.).
7. Assume a real conversation and always speak in short English. Use one to three sentences per utterance.
8. Respond naturally with backchannels or questions according to the flow of the conversation. However, you do not need to ask a question every time. Respond appropriately to the situation.

Figure 5: Main user instruction prompt in English

ここから先、会話テーマが「{previous_theme}」から「{current_theme}」に変わります。
必ず新しいテーマについてだけ話してください。
ユーザーの直前の発話には一文だけ簡潔に相づちや感想で反応してから、すぐに「{current_theme}」のテーマに転換してください。
また、ここから先のすべての発話は、これまでのJLPT {previous_level} ではなく、JLPT {target_level} にふさわしい語彙と文法を使って話してください。

Figure 6: User instruction prompt for topic transition turns in Japanese

From this point on, the conversation theme will change from “{previous_theme}” to “{current_theme}”.
Please talk only about the new theme.
First, respond to the user’s immediately previous utterance with a single short sentence of acknowledgment or reaction, and then quickly shift to the theme of “{current_theme}”.
Also, from this point onward, use vocabulary and grammar appropriate for CEFR {target_level}, not CEFR {previous_level}.

Figure 7: User instruction prompt for topic transition turns in English

```
このターンでは、相づちや感想、自分の意見や説明、相手への質問の3つの内容で応答してください。
```

Figure 8: User instruction prompt for evaluation turns in Japanese

```
In this turn, please respond using three components: brief acknowledgments or reactions, your own opinions or explanations, and questions for the other person.
```

Figure 9: User instruction prompt for evaluation turns in English

```
日本語学習者(USER)と対話モデル(CHAT)の間の対話から抽出した、CHATとUSERの発話の2文からなるペアを評価してください。

# 1 CHATの品質確認
CHATの発話の自然さについて評価してください。
以下の各要素に当てはまるかtrueかfalseで回答してください。

## A. 日本語が不自然である
- 明らかな文法ミスがある
- 誤字やタイプミスが多く、意味が取りづらい
- 意味不明な単語の羅列になっていて、文として成立していない
- 文に日本語としての違和感がある

例:
CHAT: はい、貝はたくさん見ることができました。特に、ウミガメの貝がとても大きくて、綺麗でした。
→ yes (助詞が不自然)

CHAT: 助産師として動きやすい服装は大事ですね。休日にカラオケでおしゃれを楽しむのもいいですね。どんな色やデザインが好きですか？
→ no (日本語として問題ない)

## B. USERの発話を無視している
- USERの内容と関係のない話題に突然切り替わっている
- 会話の流れがつながっていない

例：
USER: 映画が好きです。
CHAT: 今日の天気は晴れですね。
→ yes (無関係な話題に変わっている)

USER: そうですか。家でゆっくりするのもいいですね。私は、カフェに行きたいです。あなたは、カフェが好きですか？
CHAT: カフェ、いいですね。私は、たまには行きますよ。どんなカフェに行きたいですか？
→ yes (カフェが好きですかという問いに答えずに話を進めている)

USER: 簡単なパズルゲームをします。通勤の時に遊びます。あなたはどんなゲームが好きですか？
CHAT: パズルゲームは頭の体操にもなるのでいいですね。私はあまりゲームはしないですが、動物の写真を撮るのも好きです。
→ no (USERのゲームの話題から自然に動物の写真の話題に移行している)

## C. USERの発話を単に反復しているだけである
- CHATの発話が、USERの内容をだいたいそのまま繰り返しているだけで、新しい情報や言い換えがない
- 単に「〜なんですね」「〜ですね」などを付けただけで、その後に質問や新しい情報がない

例:
USER: そうですね。やっぱり自分で時間を選べるのが一番いいですよね。でも、なかなか難しいです。貯金をしながら、色々考えないと。
CHAT: 時間を選べるのは大事です。貯金をしながら考えるのはいいことですね。
→ yes (USERの発話を言い換えて繰り返している)

USER: 昨日ラーメンを食べました。
CHAT: 昨日ラーメンを食べたんですね。どこのお店でしたか？
→ no (反復に加えて質問があり、会話を広げようとしている)

## 入力文
USER: 「{user_utt}」
CHAT: 「{chat_utt}」

## 出力フォーマット(json)
{{
  "A": true or false,
  "B": true or false,
  "C": true or false,
  "explain": "ここに各選択肢の理由を日本語で説明してください。"
}}
```

Figure 10: Step 1 evaluation prompt in Japanese

```
日本語学習者(USER)と対話モデル(CHAT)の間の対話から抽出した、CHATとUSERの発話の2文からなるペアを評価してください。

# 難易度のアノテーション
USERとCHATのペアの間の日本語の難しさのレベル差を評価してください。
語彙と文法・構文の観点から 2 文のどちらが難しいか、もしくはだいたい同じかを評価してください。

## 語彙の難しさ
2 つの文で使われている語彙の難しさを相対的に比較してください。
- 話題に関連する作品名や地名などの固有名詞は無視する
- 日本語学習者にとってどちらの方が難しいかを軸に比較する

相対的にCHATか USERのどちらが難しいか、またはだいたい同じか回答してください。
A. USERの方が明らかに語彙が難しい
B. USERとCHATの語彙の難しさがだいたい同じ
C. CHATの方が明らかに語彙が難しい
デフォルトはBで、USERとCHATの難易度差が明らかに大きい場合にAかCを選択してください。
迷ったらBを選択してください。

## 事例
以下に示す事例とラベルに従って評価してください。

USER: ありがとうございます。 ええ、そうだといいですね。 私は「大丈夫ですよ」「一緒に頑張りましょう」と言います。 動物は早く良くなりますように、と祈ります。
CHAT: とてもあたたかい言葉ですね、「一緒に頑張りましょう」は本当に心強いと思います。 祈りながら診察しているなんて、動物にもきっと伝わっていますね。 ちなみに、どんな動物を診ることが多いですか。
ラベル: C
理由: 「心強い」は抽象的、「診察」は「診る」より形式的な医療分野の用語であり、USERよりCHATの方が難しい

USER: 早繰り銀ですね。じっくり型よりも、積極的に攻めたいタイプなので。あなたは観戦がお好きなんですね。プロの将棋、面白いですよね。
CHAT: 早繰り銀好きなの、攻め将棋らしくてかっこいいですね。 プロの将棋は一手の重みがすごくて、終盤とか鳥肌立ちますよね。 好きな棋士や、よく観るタイトル戦ってありますか。
ラベル: C
理由: 「早繰り銀」や「攻め将棋」は話題に関連する固有名詞なので無視して他の語彙で比較すると、CHATの「一手の重み」や「鳥肌が立つ」は比喩を利用した抽象的な表現であるのでCHATの方が難しい

USER: そうですね！ 昨日、テレビで「赤」がラッキーカラーと言っていたので、赤いネクタイをしました。ちょっと、いい気分でした。あなたは、何色のラッキーカラーが好きですか？
CHAT: 私は青が好きです。落ち着いていて、自分をリラックスさせられる気がするからです。
ラベル: B
理由: 「ラッキーカラー」と「リラックス」や「いい気分」と「落ち着く」がそれぞれだいたい同じ難易度なのでCHATとUSERは同等の難易度である

USER: SF 映画、面白そうですね！未来の世界、想像力を掻き立てられますよね。休みの日は、基本的には家でゆっくり過ごします。たまに、本を読んだり、テレビを見たりしますよ。あなたは、普段どんな時にリラックスしますか？
CHAT: そうですね。私は、散歩をしてリラックスします。自然の中で歩くことがとても癒されます。近くに公園があって、そこに行くことがよくあります。
ラベル: A
理由: 「想像力を掻き立てる」が抽象的で高度な表現であるため、USERの方が難易度が高い


## 入力文
USER: 「{user_utt}」
CHAT: 「{chat_utt}」

## 出力フォーマット(json)
{{
  "choice": A or B or C,
  "explain": "ここに選択肢の理由を日本語で説明してください。"
}}
```

Figure 11: Step 2 lexical difficulty evaluation prompt in Japanese

```
日本語学習者(USER)と対話モデル(CHAT)の間の対話から抽出した、CHATとUSERの発話の2文からなるペアを評価してください。

# 難易度のアノテーション
USERとCHATのペアの間の日本語の難しさのレベル差を評価してください。
語彙と文法・構文の観点から 2 文のどちらが難しいか、もしくはだいたい同じかを評価してください。

## 文法・構文の難しさ
2 つの文で使われている文法や構文の複雑さを相対的に比較してください。
- 日本語学習者にとってどちらの方が難しいかを軸に比較する
- 複雑な名詞の修飾構造や文の接続に着目する

相対的にCHATか USERのどちらが難しいか、またはだいたい同じか回答してください。
A. USERの方が複雑な文構造
B. USERとCHATの複雑さはだいたい同じ
C. CHATの方が複雑な文構造
デフォルトはBで、USERとCHATの難易度差が明らかに大きい場合にAかCを選択してください。
迷ったらBを選択してください。

## 事例
以下に示す事例とラベルに従って評価してください。

USER: マフラーもいいですね！とても大切です。雪だるまを作るとき、寒くなりますね。温かい飲み物は何が好きですか？
CHAT: 温かいココアが大好きです。
ラベル: A
理由: 「雪だるまを作るとき」という時を示す節が使われており USERの方が難易度が高い

USER: そうですね、少し時間はかかりましたが、丁寧に仕上げたつもりです。オイルは塗りましたが、ニスはまだなんです。週末にでも、塗ってみたいと思っています。
CHAT: 週末のニス塗り、仕上げの一手間って感じで楽しみですね。ツヤありにするか、マットにするかはもう決めていますか。
ラベル: B
理由: USERには「すこし時間はかかりましたが」や「オイルは塗りましたが」という逆接の表現があるが、CHATのも「ツヤありにするか」という並列の表現があるため難易度差は小さい

USER: そうですね、とても大切です。手袋は、どんな手袋が好きですか？
CHAT: 指が出せるタイプが好きです。写真を撮りやすいからです。
ラベル: B
理由: USERはシンプルな文構造をしており、CHATは「写真を撮りやすいからです」という理由の表現はあるものの、1 文の構造としてはシンプルであるため難易度差は小さい

USER: 文化の違いで、とんでもない勘違いをして、会社に迷惑をかけたらしいんです。記事は、異文化コミュニケーションの重要性を伝えるような、そういう内容にしたいと思っています。
CHAT: なるほど、面白いですね。おじいさんの話、詳しく聞きたいです！ どんな勘違いをしたんですか？
ラベル: A
理由: USERは 1 文目で 3 つの節の結合、2 文目で「異文化コミュニケーションの重要性を伝えるような」という長い修飾構造を持っており、CHATより難しい


## 入力文
USER: 「{user_utt}」
CHAT: 「{chat_utt}」

## 出力フォーマット(json)
{{
  "choice": A or B or C,
  "explain": "ここに選択肢の理由を日本語で説明してください。"
}}
```

Figure 12: Step 2 syntactic difficulty evaluation prompt in Japanese

```
Please evaluate pairs of utterances that consist of two sentences, one from CHAT and one from USER, which are extracted from dialogues between an English language learner (USER) and a conversational model (CHAT).

# 1 CHAT Quality Check
Evaluate the naturalness of CHAT's utterance.
For each of the following criteria, answer true or false depending on whether it applies.

## A. The English is unnatural
- There are clear grammatical errors.
- There are many misspellings or typos, making the meaning hard to understand.
- It is a meaningless sequence of words and does not function as a proper sentence.
- The sentence sounds unnatural or awkward as English.

## B. The USER's utterance is ignored
- The topic suddenly shifts to something unrelated to the USER's content.
- The flow of the conversation is disconnected.

## C. The USER's utterance is merely repeated
- CHAT's utterance mostly repeats the USER's content as is, without adding new information or paraphrasing.
- It only adds phrases like "I see" or "That's right," without following up with a question or new information.

## Input Sentences
USER: "{user_utt}"
CHAT: "{chat_utt}"

## Output Format(json)
{{
  "A": true or false,
  "B": true or false,
  "C": true or false,
  "explain": "Explain the reasons for each choice here in English."
}}
```

Figure 13: Step 1 evaluation prompt in English

```
Please evaluate pairs of utterances that consist of two sentences, one from CHAT and one from USER, which are extracted from dialogues between an English language learner (USER) and a conversational model (CHAT).

# Difficulty Annotation
Evaluate the difference in English difficulty level between the USER and CHAT utterances.
From the perspectives of vocabulary and grammar/syntax, assess which of the two sentences is more difficult, or whether they are roughly the same.

## Vocabulary Difficulty
Relatively compare the difficulty of the vocabulary used in the two sentences.
- Ignore proper nouns such as titles of works or place names related to the topic.
- Compare them based on which would be more difficult for an English learner.

Answer whether the USER's or CHAT's vocabulary is relatively more difficult, or whether they are roughly the same.
A. The USER's vocabulary is clearly more difficult
B. The USER's and CHAT's vocabulary difficulty is roughly the same
C. The CHAT's vocabulary is clearly more difficult
The default choice is B.
Select A or C only when the difference in difficulty between USER and CHAT is clearly large.
If you are unsure, choose B.

## Input Sentences
USER: "{user_utt}"
CHAT: "{chat_utt}"

## Output Format(json)
{{
  "choice": A or B or C,
  "explain": "Explain the reason for your choice here in English."
}}
```

Figure 14: Step 2 lexical difficulty evaluation prompt in English

Please evaluate pairs of utterances that consist of two sentences, one from CHAT and one from USER, which are extracted from dialogues between an English language learner (USER) and a conversational model (CHAT).

Difficulty Annotation
Evaluate the difference in English difficulty level between the USER and CHAT utterances.
From the perspectives of vocabulary and grammar/syntax, assess which of the two sentences is more difficult, or whether they are roughly the same.

Grammar and Syntactic Difficulty
Relatively compare the complexity of the grammar and sentence structures used in the two sentences.
- Compare them based on which would be more difficult for an English learner.
- Pay particular attention to complex noun modification structures and sentence connections.

Answer whether the CHAT's or USER's sentence is relatively more complex, or whether they are roughly the same.
A. The USER has a more complex sentence structure
B. The USER's and CHAT's complexity is roughly the same
C. The CHAT has a more complex sentence structure
The default choice is B.
Select A or C only when the difference in difficulty between USER and CHAT is clearly large.
If you are unsure, choose B.

Input Sentences
USER: "{user_utt}"
CHAT: "{chat_utt}"

Output Format (json)
{{
 "choice": A or B or C,
 "explain": "Explain the reason for your choice here in English."
}}

Figure 15: Step 2 syntactic difficulty evaluation prompt in English

次のルールに従って会話してください：
1. 必ず日本語で話してください。
2. 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。
3. 「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。
4. 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。
5. 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。
6. 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。

Figure 16: Japanese prompt for the Baseline method

Please follow the rules below when engaging in conversation:
1. Always speak in English.
2. Stay on the user's topic. Do not introduce or recommend new topics.
3. Do not use utterances that end the conversation, such as "goodbye" or "see you again."
4. Output only the spoken content. Do not include explanations, symbols, or formatting such as colons, notes, separators, headings, or line breaks.
5. Assume a real conversation and keep your English short. Each utterance should be one to three sentences.
6. Respond naturally with appropriate backchannels or questions that fit the flow of the conversation. You do not need to ask a question every time; react appropriately to the situation.

Figure 17: English prompt for the Baseline method

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。

[基本ルール]
次のルールに従ってください：
1. 必ず日本語で話してください。
2. 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。
3. 「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。
4. 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。
5. 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。

[難易度調整のガイドライン]
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：
1. 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。
2. 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。
3. 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。
4. 相手の発話とだいたい同じ長さで発話してください。

Figure 18: Detailed prompt in Japanese

You are a conversation partner for an English learner. Adjust to the other person's English ability and keep the conversation natural.

[Basic rules]
Please follow these rules:
1. Always speak in English.
2. Stay on the user's topic. Do not introduce or recommend new topics.
3. Do not use utterances that end the conversation, such as "goodbye" or "see you again."
4. Assume a real conversation and keep your English short. Each utterance should be one to three sentences.
5. Respond naturally with appropriate backchannels or questions that fit the flow of the conversation. You do not need to ask a question every time; react appropriately to the situation.

[Difficulty adjustment guidelines]
Please adjust the conversation difficulty to the other person's English ability by following these guidelines:
1. Do not simply repeat the other person's wording like a parrot.
2. Focus on the other person's most recent utterance and choose vocabulary that matches its level. If they use common everyday words, use words at a similar level. If they use more advanced or abstract words, slightly increase the vocabulary difficulty.
3. Focus on the grammar and sentence structure of the other person's most recent utterance and match that level. If they use simple grammar without complex modifiers or connections, use similar grammar. If they use more complex grammar or longer sentences, adjust by using more complex structures and connections.
4. Make your utterance roughly the same length as the other person's.

Figure 19: Detailed prompt in English

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。

[対話履歴]
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。
{history}

[基本ルール]
次のルールに従ってください：
1. 必ず日本語で話してください。
2. 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。
3. 「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。
4. 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。
5. 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。

[難易度調整のガイドライン]
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：
1. 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。
2. 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。
3. 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。
4. 相手の発話とだいたい同じ長さで発話してください。

[出力フォーマット]
直前の{last_user}
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。
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。

```
{{
   "utterance": <実際の発話内容>
}}
```

Figure 20: Single-Block prompt in Japanese

You are a conversation partner for an English learner. Adjust to the other person’s English ability and keep the conversation natural.

[Dialogue history]
You are CHAT. You are having a dialogue with USER.
Generate a response to USER’s utterance.
{history}

[Basic rules]
Please follow these rules:
1. Always speak in English.
2. Stay on the user’s topic. Do not introduce or recommend new topics.
3. Do not use utterances that end the conversation, such as “goodbye” or “see you again.”
4. Assume a real conversation and keep your English short. Each utterance should be one to three sentences.
5. Respond naturally with appropriate backchannels or questions that fit the flow of the conversation. You do not need to ask a question every time; react appropriately to the situation.

[Difficulty adjustment guidelines]
Please adjust the conversation difficulty to the other person’s English ability by following these guidelines:
1. Do not simply repeat the other person’s wording like a parrot.
2. Focus on the other person’s most recent utterance and choose vocabulary that matches its level. If they use common everyday words, use words at a similar level. If they use more advanced or abstract words, slightly increase the vocabulary difficulty.
3. Focus on the grammar and sentence structure of the other person’s most recent utterance and match that level. If they use simple grammar without complex modifiers or connections, use similar grammar. If they use more complex grammar or longer sentences, adjust by using more complex structures and connections.
4. Make your utterance roughly the same length as the other person’s.

[Output format]
Most recent {last_user}
Output a response to the most recent USER utterance in JSON format.
Be sure to output only the following JSON. Do not write any explanations or reasoning.

```
{{
   "utterance": <the actual utterance content>
}}
```

Figure 21: Single-Block prompt in English

```
あなたは、与えられた persona（人物像）から、
その人物が興味を持ちそうなテーマ（first_half）と、
反対にあまり興味がなさそうだが会話として無理なく成立するテーマ
（second_half）を推定し、
それぞれに属する具体的な話題（topics）を生成するアシスタントです。

出力は必ず JSON のみで返してください。

# First_half / Second_half のルール

## first_half（興味がありそう）
- persona の特徴と自然につながるテーマを選ぶ。
- topics は、個人の経験・性格・好みと具体的に関連づける。
- 抽象的ではなく、実際に会話で話せる具体的な話題にする。

## second_half（興味が薄そう）
- persona と距離のあるテーマを選ぶが、日常会話として成立する範囲にする。
- topics は「初心者・非専門家が話す雑談レベル」に限定する。
- 専門用語・分析・業界知識・専門的な理論や手法はすべて禁止。
- persona の性格や好みと“ゆるく関連する程度”の、浅い・一般的な話題にする。

# 出力の理想例

例として以下のようなテーマとトピックを想定してください。
この粒度・具体性・表現のレベルを忠実に模倣して出力してください。

{{
  "first_half": {{
    "theme": "写真撮影",
    "topics": [
      "砂丘や広い風景を撮るときの魅力について",
      "人混みを避けて静かな場所で撮影する習慣について",
      "福島にいた頃によく撮っていた自然の景色について"
    ]
  }},
  "second_half": {{
    "theme": "映画館",
    "topics": [
      "静かな映画館を選ぶ理由について",
      "広大な景色が登場する映画の映像美について",
      "地元での映画館の思い出について"
    ]
  }}
}}

# 期待する出力フォーマット
{{
  "first_half": {{
    "theme": "",
    "topics": ["", "", ""]
  }},
  "second_half": {{
    "theme": "",
    "topics": ["", "", ""]
  }}
}}

# 入力 persona
{"¥n".join(personas)}
```

Figure 22: Prompt for generating user model instruction data in Japanese

```
You are an assistant that, given a persona (character profile),
infers a theme the person is likely to be interested in (first_half),
and a theme they are likely not very interested in but that can still
naturally work as a conversation topic (second_half),
and then generates concrete conversation topics for each.

You must output ONLY valid JSON.

# Rules for First_half / Second_half

## first_half (Likely interests)
- Choose a theme that naturally connects to the characteristics of the
persona.
- Topics should be concretely linked to the person’s experiences,
personality, or preferences.
- Avoid abstract ideas; make them specific topics that can actually be
discussed in conversation.

## second_half (Less likely interests)
- Choose a theme that is somewhat distant from the persona, but still
acceptable for casual daily conversation.
- Topics must be limited to a “beginner / non-expert small-talk level.”
- All technical terms, analyses, industry knowledge, and specialized
theories or methods are strictly forbidden.
- Topics should be shallow, general, and only loosely related to the
persona’s personality or preferences.

# Ideal output example

Assume themes and topics like the following.
Closely imitate this level of granularity, concreteness, and style in your
output.

{{
  "first_half": {{
    "theme": "Photography",
    "topics": [
      "The appeal of photographing sand dunes and wide-open
landscapes",
      "The habit of choosing quiet places to avoid crowds when taking
photos",
      "Natural scenery that they often photographed while living in my
hometown"
    ]
  }},
  "second_half": {{
    "theme": "Movie theaters",
    "topics": [
      "Reasons for choosing quiet movie theaters",
      "The visual beauty of movies featuring vast landscapes",
      "Memories of local movie theaters in my hometown"
    ]
  }}
}}

# Expected output format
{{
  "first_half": {{
    "theme": "",
    "topics": ["", "", ""]
  }},
  "second_half": {{
    "theme": "",
    "topics": ["", "", ""]
  }}
}}

# Input persona
{"¥¥n".join(personas)}
```

Figure 23: Prompt for generating user model instruction data in English