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

Taku Morioka[1], Junya Takayama[1], Tomoyuki Kajiwara[1,2] (Ehime University[1], The University of Osaka[2])

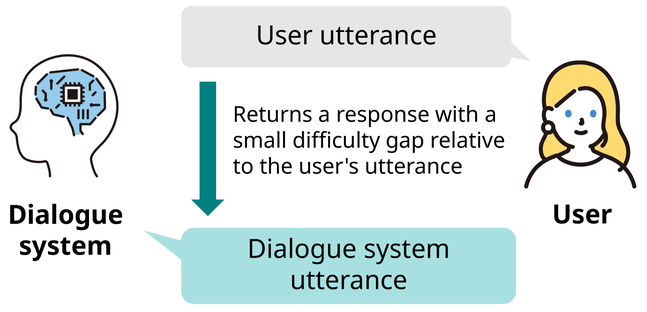
1. Background
A learner's proficiency varies with familiarity with the topic — we aim to provide utterances at an appropriate difficulty

LLM-based chatbots can offer second-language (L2) speakers a low-cost opportunity for conversation practice

Input Hypothesis[1]: obtaining comprehensible input is essential for second language learning

Following the Input Hypothesis, dialogue systems have been built that speak at a given, fixed utterance difficulty [2,3,4]

- In L2 learning, a learner's language proficiency varies with their familiarity with the topic [5]
- Prior work [2,3,4] handles fixed-difficulty instructions but does not consider dynamic difficulty changes within a dialogue



➡ **We aim to build a dialogue system that dynamically adapts to the learner's utterance difficulty**

[1] Stephen D. Krashen. (1985) The Input Hypothesis: Issues and Implications
[2] Jin et al. (arXiv-2025) Controlling Difficulty of Generated Text for AI-Assisted Language Learning
[3] Glandorf et al. (NAACL-2025) Grammar Control in Dialogue Response Generation for Language Learning Chatbots
[4] Almasi and Kristensen-McLachlan (BEA-2025) Alignment Drift in CEFR-Prompted LLMs for Interactive Spanish Tutoring
[5] Qiu (Language Teaching Research 2020) Functions of Oral Monologic Tasks: Effects of Topic Familiarity on L2 Speaking Performance

2. Proposed Method Constructing an evaluation task and difficulty-control methods

2.1 Evaluation Task
Assessing whether the system follows changes in user utterance difficulty

Step 1: Collect dialogue histories

紀伊半島の海沿いを走る予定です。いくつか聖地を巡って、温泉に入るのもいいなと思っています。あなたはツーリング好きですか？
I'm going to ride along the Kii Peninsula coast. I want to visit some special spots and try a hot spring. Do you like touring?

紀伊半島はとてもきれいな場所ですね。私もバイクで走るのが好きで、時々ツーリングに行きますよ。温泉は疲れも癒やされるので、いいですね。
The Kii Peninsula is a beautiful place. I like riding too, so I sometimes go touring. Hot springs are nice because they help you relax.

* English translations are for reference; difficulty is measured on the Japanese text.

Step 2: Evaluate the quality and difficulty gap

CHAT response quality — if any is Yes → labeled **Reject**

(a)The utterance is unnatural Yes · (No)
(b)The USER's utterance is ignored Yes · (No)
(c)The USER's utterance is merely repeated Yes · (No)

Difficulty gap between USER and CHAT

(1)Vocabulary USER harder / (Same) / CHAT harder
(2)Syntax USER harder / (Same) / CHAT harder

Final label (4 classes)

- UserHarder
- (Same)
- ChatHarder
- Reject

* A dialogue system with a higher proportion of "Same" labels is preferable

2.2 Difficulty Control Methods
Control via prompt design and fine-tuning

← Match the difficulty to the partner's utterance
↓ Methods that use an instruction prompt

Detailed method

Represent the dialogue history using the LLM chat template

```
[
  {role: system, content: "<instruction prompt>"}
  {role: user, content: "<user utterance 1>"}
  {role: assistant, content: "<chat utterance 1>"}
  {role: user, content: "<user utterance 2>"}
  {role: assistant, content: "<chat utterance 2>"}
  ...
]
```

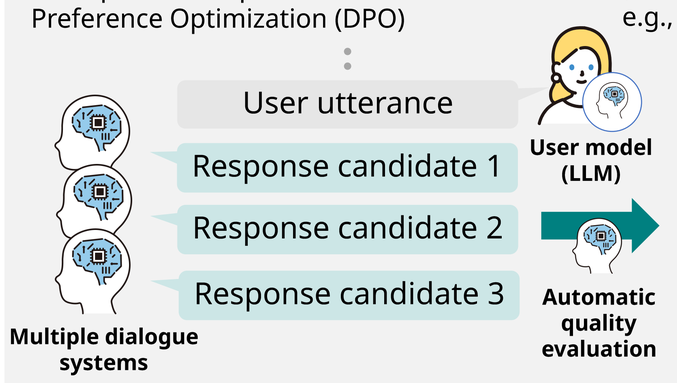
Single-Block method

Insert the dialogue history inside the prompt

```
[
  {role: user, content: "<instruction prompt> + dialogue history"}
]
```

DPO method

Build preference pairs and train with Direct Preference Optimization (DPO)



Penalty = (number of quality violations) + (number of difficulty labels that are not Same)
e.g., Yes: 2, No: 1, Same: 1, CHAT harder: 1 ⇒ 3

3. Experiments in Japanese (English results in the paper)
Single-Block and DPO are effective; human and automatic evaluation show a similar trend

From automatic evaluation (Table 1)

- Average scores rank **Baseline < Detailed < Single-Block**
  ➡ among prompt-based methods, Single-Block works best
- DPO achieves the highest score overall
  ➡ when training is feasible, DPO is the best choice

From human evaluation (Table 2)

- Baseline's ranking shifts between human and automatic evaluation
- Cohen's κ = 0.24 and Spearman's rank correlation = 0.70

➡ Human and automatic evaluation show similar trends, suggesting the validity of automatic evaluation

Experimental setup

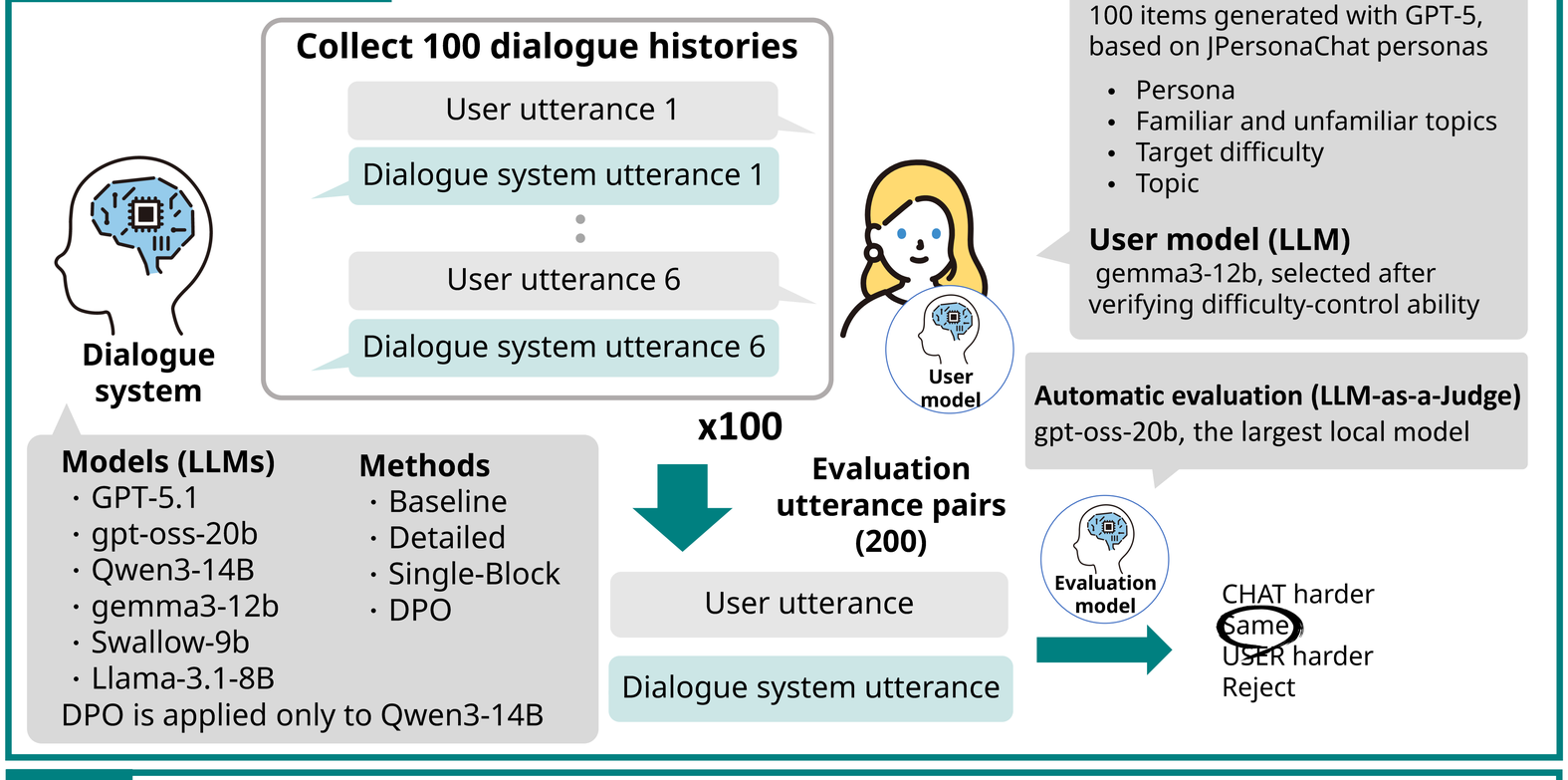


Table 1: Proportion of "Same" labels in automatic evaluation (200 utterance pairs)

Model	Baseline	Detailed	Single-Block	DPO
GPT-5.1	0.100	0.235	0.235	—
gpt-oss-20b	0.270	0.285	0.315	—
Qwen3-14B	0.205	0.280	0.320	**0.420**
gemma3-12b	**0.375**	**0.400**	**0.360**	—
Swallow-9b	0.370	0.270	0.320	—
Llama-3.1-8B	0.325	0.225	0.285	—
Avg.	0.274	0.283	**0.306**	—

Table 2: Comparison of human and automatic evaluation (100 utterance pairs)

Model	Method	Human eval.	Automatic (LLM)
GPT-5.1	Baseline	0.35	0.13
Qwen3-14B	Baseline	0.80	0.24
	Detailed	0.68	0.31
	Single-Block	0.77	0.32
	DPO	**0.85**	**0.44**

1. We propose an evaluation task for building dialogue systems that dynamically adapt to the learner's utterance difficulty
2. We suggest the validity of automatic evaluation; Single-Block and DPO each achieve high scores