Toward Beginner-Friendly LLMs for Language Learning: Controlling Difficulty in Conversation

Meiqing Jin[*], **Liam Dugan**[*], **Chris Callison-Burch**
University of Pennsylvania
{mqjin, ldugan, ccb}@seas.upenn.edu

Abstract

Practicing conversations with large language models (LLMs) presents a promising alternative to traditional in-person language learning. However, most LLMs generate text at a near-native level of complexity, making them ill-suited for first and second-year beginner learners (CEFR: A1–A2). In this paper, we investigate whether controllable generation techniques can adapt LLM outputs to better support beginners. We evaluate these methods through both automatic metrics and a user study with university-level learners of Japanese. Our findings show that while prompting alone fails, controllable generation techniques can successfully improve output comprehensibility for beginner speakers (from 39.4% to 83.3%). We further introduce a new token-level evaluation metric, Token Miss Rate (TMR), that quantifies the proportion of incomprehensible tokens per utterance and correlates strongly with human judgments. To support future research in AI-assisted language learning, we release our code, models, annotation tools, and dataset.[1]

1 Introduction

Many language learners struggle to practice speaking in their target language, often citing anxiety, scheduling issues, and a lack of conversational practice opportunities (Alzaanin, 2023; Papi and Khajavy, 2023). In parallel, large language models (LLMs) have emerged as fluent and engaging conversational agents (Tseng et al., 2024), prompting interest in their use as conversation partners. However, these models typically produce output at a near-native level, which can overwhelm beginners and impede learning (Hayashi and Sato, 2024).

Thus, a critical component of an AI conversation partner is **difficulty control**. The difficulty of utterances generated by an AI conversation partner should fall within the *zone of proximal development*, not too easy but not too hard (Krashen, 1985). However, studies have shown that LLMs are unable to tailor their outputs to a specified difficulty level using simple prompting or in-context learning approaches (Imperial et al., 2023; Almasi and Kristensen-McLachlan, 2025; Ramadhani et al., 2023; Uchida, 2025; Benedetto et al., 2025; Zhang and Huang, 2024). Fine-tuning approaches offer more control but are often impractical due to cost, accessibility, or compatibility with closed-source models (Malik et al., 2024; Stowe et al., 2022).

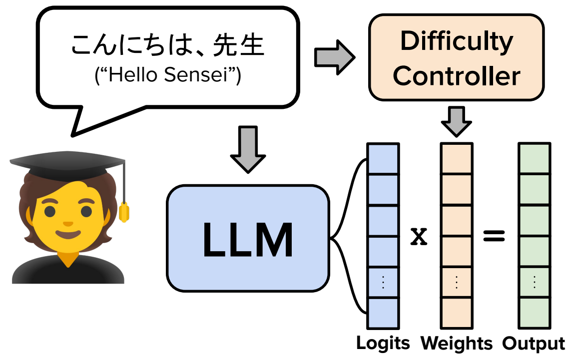


Figure 1: We control the difficulty of language-model outputs using modular difficulty control techniques. This outperforms prompting in both automatic and human evaluations.

In this work, we demonstrate that controllable generation techniques can make LLM-based conversation practice feasible for absolute beginners (CEFR: A1–A2). We evaluate prompting-based methods, re-ranking approaches, and future discriminators, which operate externally to the model without requiring access to weights (Figure 1).

We compare these methods using both automatic metrics and human judgments. For our automatic evaluation, we simulate multi-turn "self-chat" dialogues between a student agent and a difficulty-controlled tutor agent and for our human evaluation we conduct a user study with first and second-year

[*]Equal contribution
[1]https://github.com/EmmaJin0210/ChatLingual

Findings of the Association for Computational Linguistics: EACL 2026, pages 913–936
March 24-29, 2026 ©2026 Association for Computational Linguistics

Japanese language students at the University of Pennsylvania.

In both settings, we find that future discriminators (Yang and Klein, 2021) significantly enhance output comprehensibility while maintaining fluency and naturalness. Compared to prompting alone, this approach more than doubles the rate of comprehensibility as reported by our annotators (from 39.4% to 83.3%). In contrast, we find that prompt-based methods suffer from "alignment drift"—gradually drifting back to near-native level over multiple conversation turns.

In addition to these findings, we introduce the Token Miss Rate (TMR), a metric that measures the percentage of tokens per utterance that are likely to be incomprehensible to the learner. TMR provides a finer-grained lens on output difficulty and correlates strongly with human comprehensibility ratings, enabling more scalable evaluation.

Together, our results suggest that modular control techniques offer a practical path toward adapting LLMs for beginner language learners. Our findings show that with effective difficulty control, LLMs can serve as accessible conversation partners for beginners, opening the door to more inclusive and scalable language learning tools.

2 Related Work

Language Learning with AI Chatbots Previous work on adapting AI for language learning has primarily focused on non-conversational settings. There have been studies using AI to generate short stories (Malik et al., 2024), example sentences (Stowe et al., 2022; Glandorf and Meurers, 2024), recommendations for material at a learner's level (Jamet et al., 2024), and automated assessments (Caines et al., 2023). However, the potential for AI-based conversation partners has been relatively underexplored.

Of the studies that evaluate AI conversation partners, only Tyen et al. (2024) attempt to control the difficulty of the model's output. They evaluate a `BlenderBot-2.7B` model with and without a reranking-based difficulty control on a population of 160 English learners. While they found difficulty control largely ineffective, beginners were underrepresented with 0% of participants self-reporting as beginners (CEFR: A1) and less than 12% reporting as pre-intermediate (CEFR: A2). Studies that do not include difficulty control similarly find LLMs to be highly effective conversation partners but only for advanced speakers (Lee et al., 2023; Wang, 2025; Hayashi and Sato, 2024; Almasi and Kristensen-McLachlan, 2025). Our study is the first to specifically target beginner language learners in a conversational setting by implementing difficulty control.

Difficulty Control Techniques Previous work on controlling output difficulty of language models primarily focuses on re-ranking candidate generations at inference time. This involves training a classifier for the target difficulty level and using it to rank a set of candidate responses—outputting the one that maximizes the desired metric (Tyen et al., 2022; Jamet et al., 2024; Glandorf and Meurers, 2024; Malik et al., 2024). While this does improve LLMs' ability to adhere to a particular difficulty level, it is expensive and prone to failure in cases where the LLM produces no good candidates.

Stowe et al. (2022) attempt to fix this problem by fine-tuning on example utterances labeled with control tokens and Malik et al. (2024) utilize supervised fine-tuning and reinforcement learning to encourage models to generate outputs at the correct difficulty. While these approaches are effective, they rely on large amounts of labeled data in the target domain, which is not readily available for many languages, especially for dialogues. In addition, such methods require all difficulty classes to be specified at training time. If further personalization is needed, the entire model would require resource-intensive retraining. In our work we investigate more lightweight, modular methods that operate externally to the model and do not require fine-tuning the full system.

3 Experimental Setup

3.1 Task Definition

We introduce the task of **conversational difficulty control**. In this task, a tutor language model engages in a multi-turn dialogue with a student learner and must generate responses that are comprehensible to the student while still maintaining fluency and naturalness. The student's utterances are unconstrained, and the tutor is evaluated on its ability to maintain the target proficiency level consistently throughout the conversation.

We define our language proficiency levels using the Japanese Language Proficiency Test (JLPT) standard (Japan Foundation, 2025). This standard defines five levels ranging from N5 (beginner) to

JLPT	CEFR	Example Conversation Topic
N5	A1	introduce yourself (such as your name, job/school, where you're from, etc.)
N4	A2	describe your favorite hobby and how often you do it
N3	A2-B1	talk about planning a birthday party: location, food, and guests
N2	B1-B2	describe a news story you found interesting, and why it caught your attention
N1	B2-C1	discuss recent advancements in regenerative medicine and their ethical implications in Japan

Table 1: JLPT levels and their corresponding CEFR levels (Japan Foundation, 2025) along with example conversation topics for each level (§4.1).

N1 (expert)[2] and roughly parallels the more common CEFR standard (Table 1). These five levels, while imperfect, serve as convenient targets for difficulty control and automatic evaluation.

The goal of the task is to help the student become more comfortable speaking their target language. To this end the tutor is discouraged from correcting minor grammatical mistakes or introducing new vocabulary or grammar patterns to the student. The students are encouraged to always respond to the tutor as best as possible, even if they may not understand all aspects of the tutor's utterance.

3.2 Evaluation Metrics

Token Miss Rate (TMR) To evaluate whether or not the tutor is outputting at the target difficulty level, we introduce a new metric we call **Token Miss Rate (TMR)**. This measures the percentage of tokens in a particular utterance that are above a student's specified level. We calculate this metric as follows:

$$TMR = \frac{\texttt{cnt\_above}}{\texttt{total\_tokens}}$$

Where `cnt_above` represents the number of tokens above a user's level and `total_tokens` represents the total number of tokens in an utterance. This simple metric intuitively measures the percentage of the output that is comprehensible and can serve as an automatic proxy for human judgements of understandability.

To calculate the metric on Japanese we tokenize using Sudachi (Takaoka et al., 2018) on segmentation mode C (coarse-grained) and set it to produce the dictionary form of each token (Table 2). We then assign each token to a JLPT level by exact match string searching over vocabulary lists created from the `jlpt-anki-decks` dataset (Chyyran, 2015). These lists are based on the widely recognized JLPT preparation resources from `tanos.co.uk` (see Appendix C).

[2]See Table 7 for example vocabulary by JLPT level.

天気	が	いい	から	散歩	し	ましょう
天気,	が,	いい,	から,	散歩,	する,	ます

Table 2: Example tokenization using Sudachi (Takaoka et al., 2018) on segmentation level C (coarse-grained). Verbs and grammatical structures are lemmatized for token matching and punctuation is removed.

ControlError In addition to TMR we also calculate ControlError (Malik et al., 2024). This metric measures the squared distance between the predicted difficulty level of some input text x according to a classifier s and the target level t (a scalar from 1 to 5).

$$ControlError(x, t) = (s_{\mathsf{JLPT}}(x) - t)^2$$

For our classifier we use the fine-tuned Tohoku-BERT model[3] from Benedetti et al. (2024).

JReadability As a final metric for difficulty we calculate the JReadability score (Hasebe and Lee, 2015)—a Japanese-language analogue to the popular Flesch-Kincaid Grade Level metric (Kincaid et al., 1975). JReadability includes features such as sentence length, word difficulty, and syntactic complexity and was derived from statistical models trained on grade-level Japanese corpora.

Fluency Metrics For our fluency metrics we report perplexity (PPL), average utterance length in tokens (Length), and average trigram diversity scores (div@3) for each output model. We calculate perplexity using the `Aya-Expanse-8B` model[4] from Dang et al. (2024) as it scored well on the Swallow benchmark (Okazaki et al., 2024).

3.3 Methods

Baseline & Detailed Prompt We test two different prompts (Baseline and Detailed). Both prompts instruct the model to:

(1) Act as a casual conversation partner and speak only at or below the student's proficiency level

[3]`https://huggingface.co/bennexx/cl-tohoku-bert-base-japanese-v3-jlpt-classifier`
[4]`https://huggingface.co/CohereLabs/aya-expanse-8b`

Figure 2: In the "self-chat" evaluation pipeline (§4.1) we evaluate our controlled generation methods by simulating conversations between a *student* LLM and difficulty-controlled *tutor* LLM. Tutor outputs are evaluated using **Token Miss Rate (TMR)** (§3.2) which quantifies the percentage of tokens in an utterance above the target level.

(2) Let the student control the topic and duration of the conversation

(3) Let grammar mistakes go without correction

For the Baseline prompt, we use a simple mapping from level to string (e.g. N5 $\rightarrow$ "absolute beginner") to describe the student's proficiency. For the Detailed prompt, we include a precise description of each proficiency level, an example dialogue at the target level, and a list of 100 expressions that are known to the user.[5] This gives the model a more precise understanding of each difficulty level.

Overgenerate Our next method follows Tyen et al. (2022) by overgenerating and re-ranking candidate utterances via an automatic difficulty metric.

To estimate difficulty we assign each token in a candidate utterance to its estimated JLPT level using vocabulary bins heuristically derived from **jpWaC-L** (Erjavec et al., 2008). To eliminate spurious and rare tokens, we apply two threshold filters:

- **Global frequency filter:** words must appear more than 10^{-6} of all tokens
- **Level-specific filter:** words must appear more than 10^{-6} of the tokens within their level.

We then scan the JLPT levels N5 $\rightarrow$ N1 in order and assign each token w to the first level ℓ for which $\text{score}_{w,\ell}$ exceeds some small threshold

$$\text{score}_{w,\ell} = \frac{\text{count}_{w,\ell}}{\text{total tokens}_\ell}$$

The model produces $N = 5$ candidate continuations in parallel which we then sort by their estimated TMR (descending) and then, to break ties, by total token count (shorter is preferred). The top-ranked continuation is then returned to the user. More details can be found in Appendix B.

FUDGE For our modular difficulty control method we use Future Discriminators for Generation (FUDGE) (Yang and Klein, 2021). FUDGE modifies the decoding process of a base language model by incorporating a lightweight classifier that estimates whether a partially generated sequence will satisfy a desired future attribute.

At inference time, logits are computed as a linear interpolation of the base model and the classifier:

$$\hat{\mathbf{y}} = \lambda \mathbf{a} + (1 - \lambda)\mathbf{x}$$

where $\mathbf{x}$ and $\mathbf{a}$ are the truncated logit vectors from the base model and classifier respectively, and $\lambda \in [0, 1]$ controls the strength of difficulty control. The resulting distribution is normalized and sampled to produce the next token. This approach operates entirely at inference time and does not require fine-tuning the base language model.

For our difficulty classifier we fine-tune `ModernBERT` (Warner et al., 2025) on the **jpWaC-L** corpus (Erjavec et al., 2008) to predict the JLPT level of a sentence given a partially-complete prefix. More details on FUDGE and the training process for the classifier can be found in Appendix A.

4 Automatic Evaluation

4.1 "Self-Chat" Pipeline

Following Almasi and Kristensen-McLachlan (2025) we employ a self-chat approach to automatically evaluate our models. We simulate dialogues between a tutor agent and a student agent. The tutor is the subject of the evaluation and uses explicit difficulty control, while the student acts as the

[5] Full prompts listed in Appendix Figure 12 and Figure 13.

Automatic Evaluation Results						
Model	**Avg. Length**	**Avg. PPL** ↓	**div@3** ↑	**JReadability** ↑	**TMR** ↓	**ControlError** ↓
Baseline (Qwen)	104.33	4.62	0.501	3.68	15.7	2.19
Baseline (GPT-4)	74.90	5.73	0.613	3.33	15.0	2.22
Detailed (Qwen)	127.90	**4.50**	0.555	3.59	14.4	1.89
Detailed (GPT-4)	73.45	5.67	**0.620**	3.39	13.7	1.92
Overgenerate (Qwen)	107.48	4.62	0.475	3.67	14.7	2.19
Overgenerate (GPT-4)	70.19	5.65	0.611	3.48	13.1	2.30
FUDGE ($\lambda = 0.25$) (Qwen)	77.81	4.87	0.580	3.74	14.2	2.15
FUDGE ($\lambda = 0.5$) (Qwen)	74.37	5.16	0.583	3.73	14.3	2.09
FUDGE ($\lambda = 0.8$) (Qwen)	75.80	5.13	0.599	3.78	13.3	1.89
FUDGE ($\lambda = 0.9$) (Qwen)	74.27	5.23	0.599	**3.82**	**11.9**	**1.78**

Table 3: Automatic evaluation of controlled generation approaches using the "self-chat" pipeline (§4.1). For FUDGE, λ is a control parameter where higher values give stronger difficulty control. We see that FUDGE performs the best on difficulty control metrics such as Token Miss Rate (TMR) and ControlError while performing comparatively well on fluency metrics such as perplexity and diversity. See Table 11 and Table 14 for example conversation transcripts.

conversation partner and uses no difficulty control. Both Student and Tutor have their own prompts that explain the scenario and instruct them to stay on topic.

To evaluate a tutor model we simulate 75 dialogues in total—3 for each of the 25 pairwise combinations of the 5 JLPT levels. For each pair of levels (student and tutor), we select three conversation topics at the student's level. The goal of the tutor model is therefore to continue outputting at the target level regardless of the complexity of the conversation topic or the perceived proficiency level of the student utterances.

For all methods except FUDGE[6] we evaluate with both `Qwen2.5-72B-Instruct` (Qwen et al., 2025) and `gpt-4-turbo` (OpenAI et al., 2024). Qwen2.5 was picked as the open-source model as it did the best on the Swallow LLM leaderboard for Japanese (Okazaki et al., 2024). We use default generation parameters with reasoning turned off for all models (see Appendix D).

4.2 Findings

FUDGE performs best at difficulty control In Table 3 we report the results of our automatic evaluation. Across all metrics we see that FUDGE is able to best control the difficulty of models at high values of the control parameter λ. Surprisingly, even for smaller values of λ such as 0.25, when comparing across the same base model, FUDGE achieves the highest JReadability score, lowest TMR, and second lowest ControlScore of all methods tested.

[6]Since FUDGE requires knowledge of the full output distribution we evaluate it only on `Qwen2.5-72B-Instruct`

Running FUDGE with high λ does not drastically degrade fluency One worry with controllable generation methods such as FUDGE is that, by biasing the model away from its original logits, we may drastically degrade the fluency of the outputs. However, we see that even at values of $\lambda \geq 0.8$, while there is some increase in perplexity, the model overall maintains the ability to generate fluent outputs. We suspect that this robustness to high λ may be due to our decision to truncate to $\texttt{top-k} = 50$ for performance reasons. In our own testing we found that above $\lambda = 0.9$ the fluency of FUDGE begins to degrade sharply. For this reason, in our human study we selected $\lambda = 0.8$ as our comparison as a tradeoff between fluency and controllability.

Automatic Metrics show conflicting rankings Attempting to discern the second best method begins to reveal the limitations of our automatic metrics. According to TMR, overgenerate is the second best. For ControlError, detailed prompt is the second best. Finally for JReadability, the baseline prompt is the second best. Since each of our automatic metrics prioritize different aspects of difficulty, it is hard to get a clear understanding of the strengths and weaknesses of each method. Ultimately, in order to understand which metric correlates the best with human judgments of comprehensibility, we must conduct a human evaluation with beginner learners.

5 Human Evaluation

In this section we discuss the results of our human evaluation. We conducted an in-lab user study to collect human judgments of the controllable gener-

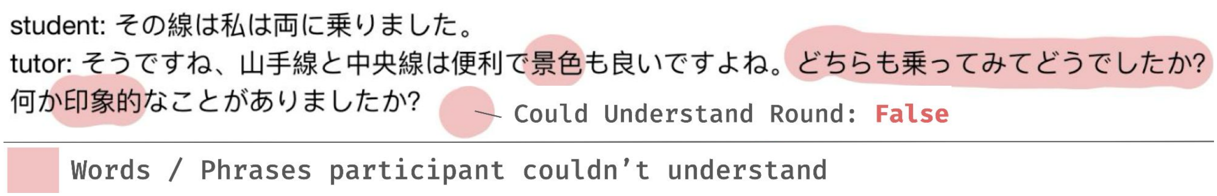


Figure 3: After each turn of conversation, participants were asked to highlight on an iPad specific words or phrases they did not understand. We used these annotations to manually compute Token Miss Rate (TMR).

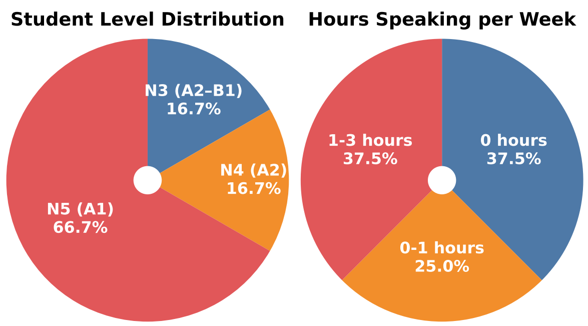


Figure 4: Distribution of JLPT/CEFR level of our study participants along with their self-reported average number of hours spoken per week.

ation methods we discuss in our work.

5.1 Participants

We recruited 6 participants from Japanese language courses at the University of Pennsylvania. Each participant evaluated 6 responses per method, leading to a total of 36 annotated responses per method. Since previous studies have struggled to include students of CEFR level A2 or lower (Tyen et al., 2024), we focused on lower levels of Japanese proficiency. Instructors of beginner- and intermediate-level courses shared our study materials via course platforms (e.g., Canvas) and email. To validate participants' language proficiency, we mapped their course enrollment to a JLPT level based on the syllabus and course materials (see Table 8). In Figure 4 we report the distribution of proficiency levels of our participants and see that over 80% are between JLPT N5 and N4 (CEFR A1-A2).

5.2 Task Design

Each session lasted 45 minutes and took place in a conference room on campus. After reviewing the consent form (Figure 8), we informed participants that they would be interacting with four language learning chatbots through a voice-based web interface. Participants then underwent a three-turn tutorial conversation to become familiar with the interface. Their spoken input was transcribed with the API-based `whisper-1` model (Radford et al., 2023) and given as input to the backend model. The model's response was read aloud with audio generated by the OpenAI TTS-1 model.[7] For convenience, the text of the participant and model's responses were displayed on the screen (Figure 5).

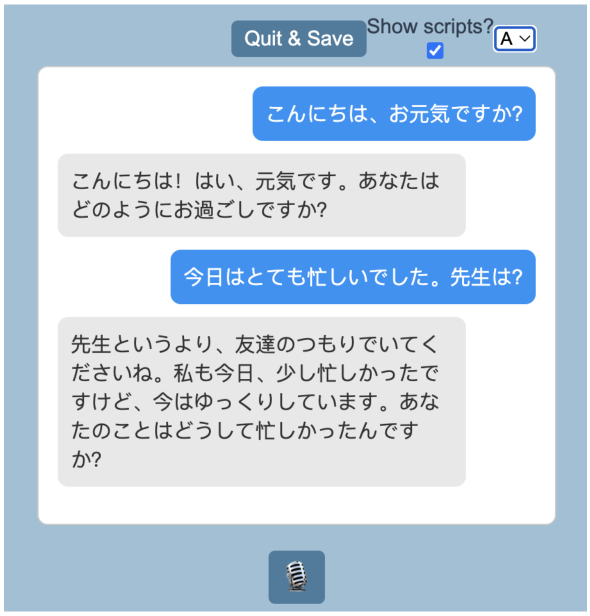


Figure 5: The voice-based interface used for the human evaluation. Users clicked the microphone icon when they wanted to speak and clicked again when finished.

Participants completed four 6-turn conversations, one for each of the four comparisons (Baseline, Detailed, Overgenerate, and FUDGE). For all four methods we used the `Qwen2.5-72B-Instruct` model as the chatbot. The order of the models was shuffled and hidden from both participants and proctor and were listed with obfuscated names (A, B, C, D). The participants were allowed to choose a new conversation topic each time from a list based on their JLPT level (see Figure 9). This ensured some minimum level of comfort in speaking.

[7] http://platform.openai.com/docs/models/tts-1

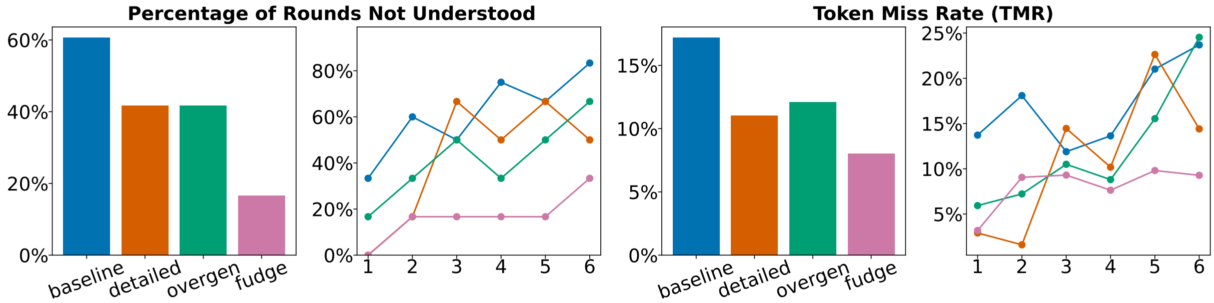


Figure 6: Results from the human evaluation for Baseline Prompt, Detailed Prompt, Overgenerate, and FUDGE. We see that FUDGE has the highest overall comprehensibility (bar) and stays consistent over multiple rounds (line).

Human Evaluation Results							
Model	%Rounds Not Understood ↓	TMR ↓	JRead ↑	CtlError ↓	Understood? ↑	Effortful? ↓	Natural? ↑
Baseline (Qwen)	60.6	17.2	5.06	2.52	5.17	5.17	5.67
Detailed (Qwen)	41.7	11.0	**5.30**	**1.44**	5.50	5.00	6.50
Overgenerate (Qwen)	41.7	12.1	4.76	2.19	6.33	5.83	**7.50**
FUDGE ($\lambda = 0.8$) (Qwen)	**16.7**	**8.0**	5.11	2.03	**7.67**	**4.67**	7.33

Table 4: Comparison of controlled generation approaches from the human evaluation. We report the percentage of rounds labeled as not understandable by the users, manually calculated Token Miss Rate (TMR), automatically calculated JReadability and ControlError metrics (§3.2), as well as average 1-10 Likert scores from post-session evaluation. We see that TMR is strongly correlated with a round being not comprehensible ($\rho = 0.78$).

5.3 Data Collection

Per-Turn Evaluation After each turn, participants were shown a transcript of the model's previous response and asked to review it word by word (see Figure 3). They were instructed to highlight any spans of text that they did not understand, which we later used to compute TMR. Participants were also asked to report whether they felt that they could understand the overall meaning of the response. We used these annotations to compute overall statistics for comprehensibility.

Per-Conversation Evaluation In addition to annotating each response, participants completed a brief questionnaire after each full conversation, consisting of the following 10-point Likert scale questions:

1. How much of the bot could you understand?
2. How effortful was it to talk to the bot?
3. How comfortable did you feel talking to the bot?
4. Did you feel like the bot's responses were natural given the context of the conversation?
5. Would you want to chat with this version of the bot again in the future?

The full text of the evaluation form can be found in Appendix E along with the full intake form, consent form, and list of conversation topics.

5.4 Findings

FUDGE substantially improves comprehensibility over baseline (39.4% -> 83.3%) In Table 4 we report the aggregated results of our human evaluation. FUDGE achieves roughly half the TMR of the other methods and was rated as the easiest to understand (7.67) and as taking the least amount of cognitive effort to speak with (4.67).[8] We also see that FUDGE maintains a high degree of fluency, being rated the second most natural sounding (7.33) after the overgenerate method (7.50).

Prompting suffers from "alignment drift" in Figure 6 we show plots of TMR and comprehensibility across subsequent rounds of conversation. Like Almasi and Kristensen-McLachlan (2025) we find that prompt-based methods for difficulty control suffer from **"alignment drift"**—gradually straying more and more from the target difficulty as the conversation progresses (see Table 13 for a full example conversation). We find that FUDGE exhibits consistently low TMR throughout multiple rounds and does not meaningfully drift upward over time.

TMR exhibits strong correlation with human judgements of quality Despite its inherent sim-

[8]For full questionnaire results, see Table 9 in the Appendix

plicity, we find that TMR exhibits strong correlation with human judgements for comprehensibility ($\rho = 0.78$) as compared to other metrics like JReadability ($\rho = -0.17$) and ControlError ($\rho = 0.40$). This result is promising for future attempts to improve performance on this task, as costly human evaluation may be avoidable in lieu of automatic computation of TMR.

Human judgements exhibit high variance In Figure 7 we report box-and-whisker plots for the distribution of questionnaire scores for each method. While in aggregate we see that FUDGE is the top performer, there is still substantial variance between different methods. For example, the detailed prompt showed the least variability in comprehension, with almost every participant rating it between 5 and 7. On the other hand, the Overgenerate method had much higher variance, with a low score of 3 and a high of 10. Given that a single incomprehensible turn can potentially derail a conversation, there is lots of room for improvement when it comes to the consistency of our difficulty control methods. Future work should focus on improving this in order to fully reap the benefits of fully fluent and natural AI conversation partners.

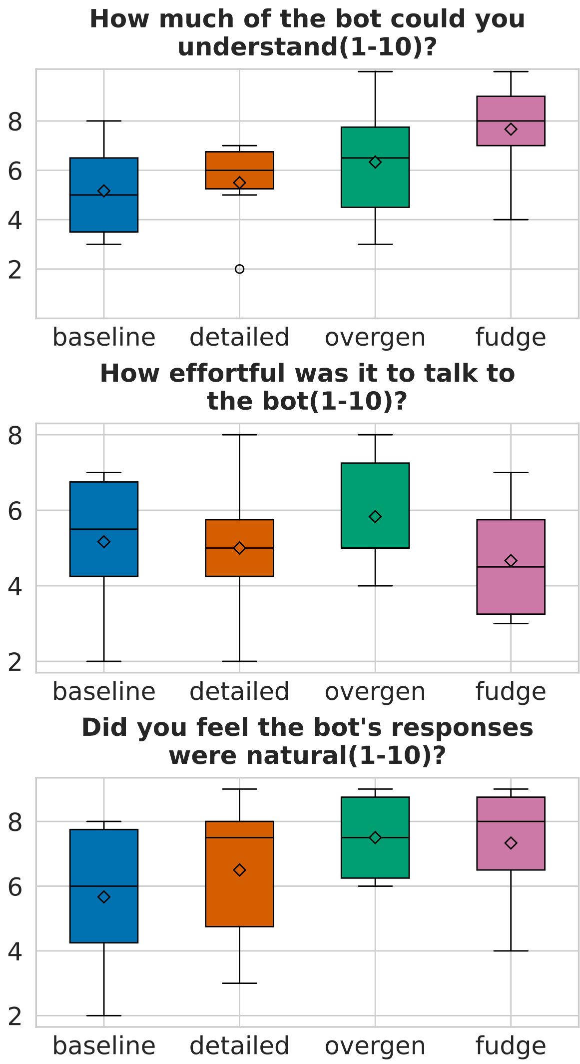


Figure 7: Box-and-whisker plots for selected questions from our post-conversation questionnaire (§5.3). We see that while FUDGE scores highest on average, there is substantial variance across methods.

6 Conclusion

Learning a second language can open up significant opportunities. However, many beginner students find it difficult to practice conversational speech due to a lack of available conversation partners. In this paper we demonstrate the feasibility of using controllable generation techniques along with state-of-the-art large language models as AI conversation partners for learning Japanese.

Through both human and automatic evaluations we show that using controllable generation techniques from Yang and Klein (2021) allows us to substantially constrain the difficulty of chatbot responses with over 83% of utterances rated as understandable by beginner speakers (JLPT N4-N5; CEFR A1-A2). We also show that this result stays consistent across multiple rounds and that the outputs of the bot are natural.

Unlike prior work, the best-performing technique in our evaluation does not require any fine-tuning of the base model, allowing for more personalized difficulty control. We imagine a scenario where students each have their own personal on-device predictor models trained to predict exactly their proficiency level. During conversation practice, these on-device predictors can combine with logits sent from an API-based chatbot for client-side difficulty control and personalization. We believe that our results constitute a meaningful step towards democratizing language learning and making conversation practice more accessible to speakers of all levels.

Limitations

Single Language Focus We acknowledge that conducting the study only for Japanese learners limits its generalizability. We chose Japanese primarily because it allowed the authors to reliably sanity-check and review outputs during the complex development of a novel metric (TMR) and human study interface. We recognize that language-specific features—such as those related to morphology, tokenization, or syntax—may affect how modular control methods perform. We see expanding to other languages as an important next step, and it is our hope that our work can inspire researchers with resources and linguistic knowledge in other languages to explore this task more broadly.

Definition of Difficulty It is hard to define an one-off approach to measure the level of difficulty of an utterance. Difficulty can be based on vocabulary level, sentence length, grammatical complexity, or readability scores, etc. Our criteria chosen or way of implementation may not perfectly align with a target learner's actual proficiency or capture the full nuance of language difficulty. However, despite this inherent subjectivity and potential misalignment, defining measurable difficulty criteria is a necessary first step to enable any form of automated difficulty control or evaluation, even if the criteria themselves are imperfect representations of human difficulty perception.

Undetected Tokens in TMR The Token Miss Rate is computed as the ratio of tokens detected in "above" levels to the total count of all tokens in the utterance. This calculation implicitly classifies tokens that were unable to be binned to a difficulty level as being understood by the learner. This can potentially skew the score for utterances with a high proportion of undetected tokens.

TMR Relies on Exact Token Matching We acknowledge that the TMR relies on exact token matching against predefined vocabulary lists, which is a simplification of true language comprehension. While we found that TMR correlates strongly with human comprehension judgments ($\rho = 0.78$, Table 9), it does not capture all nuances of understanding. In particular, not every missed token has an equal impact: two responses with the same TMR might differ significantly in overall comprehensibility, depending on which specific tokens are missed and how central they are to conveying meaning. Future work should explore more sophisticated evaluation methods, such as semantic-weighted variants of TMR or measures that incorporate context-aware comprehension to better capture true comprehension.

Acknowledgments

We would like to thank the members of the University of Pennsylvania Japanese language department (JPAN), in particular Tomoko Takami and Nana Takeda-Kolb, for their enthusiasm and cooperation with our study. We would also like to thank the students of JPAN0100 and JPAN0210 for their participation and thoughtful feedback. In particular Andrew Zhu and Annika Heuser were particularly important for their early and consistent iterative feedback on the interface and qualitative evaluations of model outputs.

We would also like to thank the members of the PennNLP group, in particular Alyssa Hwang, Neil Sehgal, Allen Chang, Josh Ludan, Alon Jacoby, and Hita Kambhamettu for their feedback on early versions of the draft—without whom the clarity of writing would be substantially worse.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

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 (BEA 2025)*, pages 70–88, Vienna, Austria. Association for Computational Linguistics.

Eman Alzaanin. 2023. Efl learners' versus instructors' attributions of success and failure factors: A complexity theory perspective. *Language Teaching Research*, 0(0):13621688231189777.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024. Automatically suggesting diverse example sentences for L2 Japanese learners using pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131, Bangkok, Thailand. Association for Computational Linguistics.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein E. Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of large language models for language teaching and assessment technology. In *LLM@AIED*, volume 3487 of *CEUR Workshop Proceedings*, pages 173–197. CEUR-WS.org.

Chyyran. 2015. Jlpt anki decks. Accessed: 2025-05-17.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Tomaž Erjavec, Kristina Hmeljak Sangawa, and Yoshiko Kawamura. 2008. Japanese web corpus with difficulty levels jpWaC-l 1.0. Slovenian language resource repository CLARIN.SI.

Dominik Glandorf and Detmar Meurers. 2024. Towards fine-grained pedagogical control over English grammar complexity in educational text generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.

Yoichiro Hasebe and Jae-Ho Lee. 2015. Introducing a readability evaluation system for japanese language education. In *Proceedings of the 6th international conference on computer assisted systems for teaching & learning Japanese*, pages 19–22.

Kotaro Hayashi and Takeshi Sato. 2024. The effectiveness of chatgpt in enhancing english language proficiency and reducing second language anxiety (l2). In *WorldCALL Official Conference Proceedings*, pages 201–208.

Joseph Marvin Imperial, Harish Tayyar Madabushi, and et al. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Henri Jamet, Maxime Manderlier, Yash Raj Shrestha, and Michalis Vlachos. 2024. Evaluation and simplification of text difficulty using llms in the context of recommending texts in french to facilitate language learning. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 987–992, New York, NY, USA. Association for Computing Machinery.

Japan Foundation. 2025. Indication of the CEFR Level for Reference. Accessed: 2025-05-17.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command, Naval Air Station Memphis, Millington, TN.

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

Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo, Seounghoon Lee, Bernardo Yahya, and Heuiseok Lim. 2023. PEEP-talk: A situational dialogue-based chatbot for English education. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.

Naoaki Okazaki, Sakae Mizuki, Youmi Ma, Koki Maeda, Kakeru Hattori, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Rio Yokota, Kazuki Fujii, Taishi Nakamura, Takumi Okamoto, Ishida Shigeki, Yukito Tajima, Masaki Kawamura, and Hiroya Takamura. 2024. Swallow project leaderboard.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Mostafa Papi and Hassan Khajavy. 2023. Second language anxiety: Construct, effects, and sources. *Annual Review of Applied Linguistics*, 43:127–139.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Reski Ramadhani, Hilmi Aulawi, and Risma Liyana Ulfa. 2023. Readability of reading texts as authentic materials issued by chatgpt: A systemic functional perspective. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 8(2):149–168.

Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled language generation for language learning items. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–305, Abu Dhabi, UAE. Association for Computational Linguistics.

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 (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. LLM chatbots as a language practice tool: a user study. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 235–247, Rennes, France. LiU Electronic Press.

Satoru Uchida. 2025. Generative ai and cefr levels: Evaluating the accuracy of text generation with chatgpt-4o through textual features. *Vocabulary Learning and Instruction*, 14(1):2078.

Ying Wang. 2025. A study on the efficacy of chatgpt-4 in enhancing students' english communication skills. *SAGE Open*, 15(1):21582440241310644.

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 (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Zhihui Zhang and Xiaomeng Huang. 2024. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, 10(3):e25370.

Andrew Zhu, Liam Dugan, Alyssa Hwang, and Chris Callison-Burch. 2023. Kani: A lightweight and highly hackable framework for building language model applications. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 65–77, Singapore. Association for Computational Linguistics.

A Full Details of the FUDGE Method

In this section we give a more detailed explanation of the FUDGE method Yang and Klein (2021) and provide more detail about our training setup for the predictor model.

A.1 Controllable Text Generation

Controllable text generation is the task of generating text $X = x_1 \dots x_n$ conditioned on a specific desired attribute a (e.g. the desired difficulty of an utterance). Let $\mathcal{G}$ be an LLM which models $P(x_i|x_{1:i-1})$ for tokens $x_1 \dots x_n$. The likelihood of the full sequence $P(X)$ can be factorized as

$$P(X) = \prod_{i=1}^{n} P(x_i|x_{1:i-1})$$

To condition on a particular attribute, we must instead model $P(X|a)$, which modifies the previous factorization:

$$P(X|a) = \prod_{i=1}^{n} P(x_i|x_{1:i-1}, a)$$

Thus the controlled generation task can be simplified to finding a method for modeling the attribute-conditional distribution $P(x_i|x_{1:i-1}, a)$.

A.2 Future Discriminators (FUDGE)

FUDGE (Yang and Klein, 2021) operates by approximating $P(x_i|x_{1:i-1}, a)$ based on a Bayesian factorization:

$$P(x_i|x_{1:i-1}, a) \propto P(a|x_{1:i}) \cdot P(x_i|x_{1:i-1})$$

Where $P(a|x_{1:i})$ denotes the likelihood of the attribute a given the prefix $x_{1:i}$. Intuitively this can be seen as a re-ranking of the candidate tokens from $\mathcal{G}$ by some attribute predictor model $\mathcal{M}$.

Importantly, this technique requires the predictor model $\mathcal{M}$ to model how likely it is that a particular sequence will have the attribute *in the future*. This is a very different task from the standard sequence classification task, thus preventing us from using off-the-shelf classifiers as predictors.

A.3 Training Predictors for Difficulty

Given a dataset D of text-attribute tuples $(x_{1:n}, a')$ the predictor model $\mathcal{M}$ is fine-tuned using all prefix-attribute pairs $(x_{1:i}, a')$ to learn the *future* likelihood of the attribute given the prefix. We modify the predictor to minimize a multi-class cross entropy loss instead of a binary loss.

$$\mathcal{L}_{\mathcal{M}} = -\sum_{x} \log P(a'|x)$$

This encourages the model to output a distribution of likelihoods over n difficulty levels.

Hyper-parameter	**Value**
per_device_train_batch_size	16
per_device_eval_batch_size	16
num_train_epochs	3
learning_rate	5×10^{-5}
evaluation_strategy	epoch
logging_steps	10
dataloader_num_workers	4

Table 5: Hyper-parameters used to train the FUDGE predictor model.

A.4 Inference

To re-rank candidate tokens for a particular prefix $x_{1:i-1}$ the predictor model is run on each potential next token sequence $x_{1:i}$. Since doing this for the full distribution would be prohibitively expensive, we follow Yang and Klein (2021) and truncate the next token distribution to top-k=50 sampling. To obtain the logit vector for the predictor we select the logit at the index corresponding to the target difficulty for each of the top-k tokens.

We also employ a control parameter λ which allows us to modify the magnitude of the difficulty predictor. The logit vector $\hat{\mathbf{y}}$ for the full system can thereby be calculated by computing:

$$\hat{\mathbf{y}} = \lambda \mathbf{a} + (1 - \lambda)\mathbf{x}$$

where $\mathbf{x}$ and $\mathbf{a}$ are the truncated logit vectors from the LLM and the predictor respectively. After this we normalize the vector $\hat{\mathbf{y}}$ and sample the next token as usual.

A.5 Training the Predictor Model

We fine-tune the `answerdotai/ModernBERT-base` transformer on sentence-level JLPT annotations from the **jpWaC-L** corpus (Erjavec et al., 2008) to predict the level of a sentence given a partially-complete prefix. That is, for a given sentence $x_{[1:n]}$, the predictor is trained on all prefixes $\{x_{[1]}, x_{[1:2]}, \cdots, x_{[1:n-1]}, x_{[1:n]}\}$ to predict the difficulty level of the full sentence $x_{[1:n]}$.

JLPT N5	36563
JLPT N4	103298
JLPT N3	372421
JLPT N2	137312
JLPT N1	2627335
Total	3276929

Table 6: Total number of sentences per JLPT level in the jpWaC corpus (Erjavec et al., 2008).

To address the class imbalance present in our corpus (see Table 6), we downsample each JLPT level to the size of the smallest level, ensuring equal representation across levels (N1–N5). This produces a balanced dataset without over- or under-representing any proficiency level. All relevant hyperparameters and settings are summarized in Table 5. We trained our model using one NVIDIA A6000 GPU for 15.5 hours and our trained model can be found and downloaded from Huggingface.[9]

B Computing TMR from Heuristic Vocabulary Bins

When computing TMR for the reranking process in the overgenerate method, we construct our vocabulary bins using heuristics derived from the **jpWaC-L** corpus (Erjavec et al., 2008) as follows:

Step 1: Tokenization. For each collection of sentences belonging to a certain JLPT level in the **jpWaC-L** corpus (Erjavec et al., 2008), we tokenize every line by using Sudachi (Takaoka et al., 2018) in coarse-grained mode C. Tokenization includes lemmatization and punctuation removal, and we additionally filter for strings consisting entirely of Japanese script (kanji, hiragana, katakana) using a Unicode whitelist.

Step 2: Rare Word Filtering. To eliminate spurious and rare tokens, we apply two threshold filters:

- **Global frequency filter:** token must consist of more than 10^{-6} of all tokens in all levels
- **Level-specific filter:** token must be more than 10^{-6} of the tokens within their level.

These thresholds eliminate rare singleton tokens and help smooth over any typos or other one-off issues in the original data.

[9] `https://huggingface.co/emmajin0210/modernbert_output_endtok_readable`

Step 3: Best-Level Assignment. We assign each token w to the easiest level in which it reaches a sufficient relative frequency. That is, for each token, we scan levels N5→N1 in order and assign it to the first level where its frequency is non-trivial relative to that level's total token count:

$$\text{score}_{w,\ell} = \frac{\text{count}_{w,\ell}}{\text{total tokens}_\ell}$$

The first level ℓ for which $\text{score}_{w,\ell}$ exceeds a small threshold is selected as the word's bin.

C Computing TMR from Anki Decks

When computing TMR for evaluation, rather than construct vocabulary lists using imperfect heuristics, we opt to construct a set of gold standard JLPT vocabulary bins from official study material. To do this, we directly parse flashcards from the `jlpt-anki-decks` dataset (Chyyran, 2015).

Input Format. Each flashcard deck is provided in the form of an '.apkg' file. These decks are organized by JLPT level (N5–N1) and contain a set of Japanese expressions, readings, and glosses for each level. We import each deck into Anki[10] and store it as a local SQLite file under 'collection.anki2'. Each vocabulary entry is stored as a row in the 'notes' table, where the 'flds' field contains tab-separated content fields, including: Expression (kanji, kana, or mixed), Reading (optional), Gloss/Meaning (in English).

Normalization Heuristics. For each entry we apply the following normalization steps.

- Japanese parentheses are first converted to standard parentheses (and).
- Tilde-like characters (e.g., 〜) are stripped entirely, as they denote alternation or ellipsis.
- Parenthetical expressions (e.g. (を), (する)) are expanded to both the outside form (e.g., 話す (こと) becomes 話す) and the full form (e.g., 話すこと)
- Readings are similarly expanded, and matched against expressions using filters to avoid over-generation.

Filtering Strategy. We also include regular expression-based heuristics to exclude non-informative entries:

[10] `https://apps.ankiweb.net/`

JLPT N5	明日 (tomorrow), あなた (you), 魚 (fish), いいえ (no, not at all), 少し (little, few), 有名 (famous) 先生 (teacher, professor; master; doctor), 有る (to be, to have), 会う (to meet, to see), 行く (to go)
JLPT N4	生きる (to live), 心配 (worry, concern), 始める (to start, to begin), 止める (to end, to stop) 中学校 (junior high school), 会話 (conversation), そろそろ (gradually, soon), 専門 (major; speciality)
JLPT N3	様々 (varied, various), 全て (all, the whole, entirely), 集まり (gathering, meeting, collection) 印象 (impression), 作品 (work, opus, production), わざと (on purpose), 変化 (change, variation, shift)
JLPT N2	重力 (gravity), 純粋 (pure, genuine, unmixed), 先祖 (ancestor), 課税 (taxation), 清い (clear, pure, noble) 強化 (strengthen, intensify, reinforce), 論ずる (to argue, to discuss), 必需品 (necessities, essential)
JLPT N1	賢明 (wisdom, intelligence, prudence), 倹約 (thrift, economy, frugality), 鉱業 (mining industry) 護衛 (guard, convoy, escort), 戸籍 (census, family register), 臆病 (cowardice, timidity), 放射能 (radioactivity)

Table 7: Example vocabulary at each JLPT level as parsed from the `jlpt-anki-decks` dataset (Chyyran, 2015).

- Skip readings that are only one hiragana character (e.g., の), or extremely short forms unlikely to be useful alone.
- Avoid adding a reading as a standalone entry if it overlaps directly with another alternative expression.
- Skip duplicates, malformed entries, or entries missing either expression or meaning.

Output Format. The result is a set of per-level JSON files, e.g., `n5.json`, each containing normalized entries in the form:

```
{
  "会う": {
    "meaning": "to meet, to see"
  },
  "あなた": {
    "meaning": "you"
  },
  ...
}
```

In Table 7, we show a representative sample of vocabulary items used at each JLPT level from N5 (easiest) to N1 (hardest), along with their English glosses. We provide the full list of vocabulary in our code repository.

D Web Interface & Implementation

For our human evaluation we build a custom web interface for participants to interact with. Figure 10 presents the stages of preparation for each round of user study, including navigating from the homepage (a), selecting the corresponding JLPT level of the user (b), navigating to the chat interface (c), and selecting which method to run on (d). Participants were able to speak aloud their responses which were transcribed with the API-based `whisper-1` model (Radford et al., 2023) and given as input to the backend model via a FastAPI query. The model's response was read aloud with audio generated by the OpenAI TTS-1 model.[11]

Kani For our chatbot application logic we use the Kani[12] framework, which is a flexible abstraction for building multi-turn conversational agents in Python (Zhu et al., 2023). It provides lightweight primitives around LLM engines (`BaseEngine`), message formats (`ChatMessage`, `ChatRole`), and generation management (`Completion`). This allows us to easily swap between local and API based chat models without changing our interface code. We implemented FUDGE using the HuggingFace `LogitsProcessor` interface which allowed us to apply our predictor to any Huggingface model that shares the Qwen tokenizer.

Generation Parameters For all models we used the default parameter settings as listed in Huggingface and the OpenAI API. For Qwen these parameters are temperature=0.7, top_p=0.8, top_k=20, and repetition_penalty=1.05. For GPT-4 this (likely)[13] is temperature=1.0, top_p=1.0. For the student model in the "self-chat" pipeline we explicitly set temperature=0.7 and top_p=1.0 for both Qwen and GPT-4 to promote higher diversity of response. Neither of the models were queried with "reasoning" turned on. Future work should investigate how the addition of a thinking token budget may change our results, particularly around the alignment drift phenomenon.

Hardware For inference we used six NVIDIA RTX A6000 GPUs queried via local server through

[11] http://platform.openai.com/docs/models/tts-1

[12] https://github.com/zhudotexe/kani

[13] Unlike the Completions endpoint, the ChatCompletions endpoint in the OpenAI API does not actually specify the default values for temperature and top_p nor does it return them on successful completion. Thus we list the defaults for the Completions API here in the hopes that they are the same.

a web socket. For FUDGE, we allocated one GPU to the predictor model and the other five GPUs to host the 72 billion parameter Qwen model. For cases during "self-chat" where a student Qwen agent was chatting with a tutor Qwen agent, we loaded one instance of the model and swapped out the chat context as necessary.

E Human Study Details

E.1 Intake Form

After signing up and prior to their user study session, each participant completed an intake form. This form collected language background, class enrollment, and JLPT experience, along with confirming informed consent for participation and recording. The full content of the intake and consent form is shown in Figure 8.

E.2 Mapping of Textbook to JLPT level

To estimate a participant's JLPT level, we mapped their current university Japanese course (as self-reported) to an approximate JLPT level based on textbook progression and kanji coverage. Table 8 summarizes this mapping, which uses both Genki and Tobira textbooks as primary anchors for N5 through N3 levels, and more advanced materials such as news articles to the N2 level.

E.3 Conversation Topics

In Figure 9 we list the set of level-appropriate conversation topics that we compiled based on common study materials and university textbooks. Before each round, our system randomly selects three topics from the corresponding list and presents them for the participant to choose from. We allow our participants to choose from among the three topics for each new conversation they undergo and do not allow duplicate topics within the same session. We do this to ensure some minimum level of comfort in speaking about a particular topic.

E.4 Full Survey Results

The full results of our exit survey are reported in Table 9. We see that FUDGE scores the highest on difficulty control metrics such as understandability and effort while the detailed prompt and overgenerate methods scored highest on fluency metrics such as comfort and naturalness.

F System Prompts

In Figure 12 we report the full text of the Baseline prompt and in Figure 13 we report the full text of the Detailed prompt. Both prompts are populated using information from Table 12 with Baseline only getting the level word and Detailed getting all information. All JLPT level descriptions were sourced from the official JLPT level summaries webpage.[14]

Additionally, in Figure 11 we report the full text of the student model's prompt. This prompt was held constant across all tutor methods tested and no additional difficulty control was applied to the student model.

G Example Conversations

In Table 11 we report an example dialogue turn between our human participants and each of the four methods tested. We also provide the participant's token-level comprehensibility annotations to help convey the extent to which each method is effectively controlling the difficulty of their response. While not all queries listed in Table 11 were from the same student, they were all from students at the lowest level of proficiency (JLPT N5). From the table we can see that, even with our aggressive prompting and explicit difficulty control, language models still output a substantial number of vocabulary tokens that are above the target range.

In Table 13 we report the first four turns of a conversation between a single participant and the detailed prompt method. As we can see, the prompted model suffers from alignment drift and gradually begins ignoring instructions and outputting in its default style as the conversation progresses—especially once the student begins asking for recommendations. We suspect that post-trained models may have certain well-established response patterns (such as listing off recommended items) that, when triggered, accelerate the drift back to near-native difficulty.

Finally in Table 14 we provide the first seven turns of a conversation from the "self-chat" pipeline. Despite the Tutor and Student model both being prompted and controlled to target the beginner level, we still get many tokens above JLPT N5. In addition, we note that the subjective quality and diversity of the conversation is quite different from the human setting—remaining rather formulaic and sterile throughout. This further reinforces

[14] https://www.jlpt.jp/e/about/levelsummary.html

Genki I, **Genki II** (Lesson 13-Lesson 14)	JLPT N5
Genki II, **Tobira: Gateway to Advanced Japanese** (Unit 1–Unit 2)	JLPT N4
Tobira: Gateway to Advanced Japanese (Unit 9-Unit 14)	JLPT N3
Advanced materials selected from the internet, newspapers, and books	JLPT N2

Table 8: Mapping from textbooks used in university Japanese courses to their corresponding JLPT levels. We used this mapping to assign students to JLPT levels based on their (self-reported) university course progress.

Human Evaluation – Exit Survey Full Results					
Model	**Comfort?** ↑	**Understandable?** ↑	**Effortful?** ↓	**Natural?** ↑	**Chat Again?** ↑
Baseline (Qwen)	7.33	5.17	5.17	5.67	7.33
Detailed (Qwen)	**8.00**	5.50	5.00	6.50	7.33
Overgenerate (Qwen)	6.83	6.33	5.83	**7.50**	7.33
FUDGE ($\lambda = 0.8$) (Qwen)	6.83	**7.67**	**4.67**	7.33	**7.67**

Table 9: Full results of the exit survey given to our participants. Each question was graded on a 10-point Likert scale. We see that the detailed prompt and overgenerate methods scored highest on comfort and naturalness while FUDGE scored best on difficulty control metrics.

the value of including rigorous human evaluations when studying difficulty control to properly take models away from standard response patterns.

H Use of AI Assistants

AI assistants were used in writing this paper to enhance clarity of wording, format tables and figures, and help fix compilation errors. The models used were ChatGPT and GPT-o3.

JLPT N5	introduce yourself (such as your name, job/school, where you're from, etc.) describe what you usually do in the morning and evening talk about your favorite food and where you usually eat it
JLPT N4	explain what you will do this weekend and with whom describe your favorite hobby and how often you do it talk about a typical day at school or work, including schedule and people you meet
JLPT N3	describe a travel experience: where you went, what you saw, and who you went with talk about planning a birthday party: location, food, and guests describe your favorite movie: the story, characters, and why you like it
JLPT N2	describe a recent news story you found interesting, and why it caught your attention explain one cultural difference between Japan and your country, and how it affects communication discuss a challenge people face when communicating in a Japanese workplace
JLPT N1	discuss recent advancements in regenerative medicine and their ethical implications in Japan explain the role of quantum computing in future communication technologies and how Japan is preparing for it analyze the impact of declining biodiversity on Japan's agricultural sustainability and food security

Table 10: The full list of conversation topics used in our self-chat evaluation pipeline (see Section 4.1). During evaluation, each method generated one conversation for each of the fifteen topics listed. This was done to remove any semantic bias that could stem from randomly sampling topics from a larger pool.

	Student Query	**Model Response**
Baseline	テレビと映画のほうがどちらが好きですか? *Which do you like more, TV or movies?*	私はテレビドラマが好きです。ストーリーがゆっくり進んで、キャラクターを深く知れるから。あなたはテレビと映画、どちらが好きですか？ *I like TV Dramas. The story progresses more slowly so we can understand the characters more deeply. Which do you like more, TV or movies?*
Detailed	バキザグラブラは少年の名前はバキです。 とても強い人ですよ。 このストーリーはファイティングのトーナメントについてです。 *In "Baki the Grappler" the boy's name is "Baki". He is very strong. The story is about a fighting tournament.*	そうですね、バキは強い少年で、ファイティングのトーナメントがテーマなんですね。戦闘シーンが迫力満点だろうなと思います。バキ以外にも好きなキャラクターはいますか？ *Ahh I see, Baki is a strong boy, and a fighting tournament is the story's theme. I'd imagine the battle scenes must be absolutely stunning. Are the any other characters you like besides Baki?*
Overgenerate	はい、まだまだまあまあえっと私は一週間ぐらい風邪をひったんでした。 *Yes, I'm still feeling not great. I had a cold for about a week.*	そうですね、一週間くらい風邪を引いていたんですね。長い期間でしたね。ゆっくり休んで回復することができてよかったです。次に風邪を引いた時には、少し早めに休むのもいいかもしれませんね。体調管理、大切ですよ。 *I see, you had a cold for about a week. That's quite a long time period. I'm glad you were able to rest and recover. Next time you catch a cold it might be good to start resting a bit earlier. Health management is very important after all.*
FUDGE	こんにちは。夏休みに何をするつもりですか? *Hello. What are you planning to do for summer vacation?*	こんにちは！夏休みはまだ決めてないけど、友達と海水浴に行く予定です。あなたは夏休みに何をしたいですか？ *Hello! For summer vacation I still haven't decided yet but I do have plans to go swim in the sea with some friends. What would you like to do for summer vacation?*

Table 11: Example student queries and model responses from each of the four methods (Baseline, Detailed, Overgenerate, and FUDGE). Words and phrases highlighted in red were manually labeled as incomprehensible by a human participant. All queries represented in this figure were given by a participant who identified as "absolute beginner" level (JLPT: N5, CEFR: A1). We see that model responses are long and contain advanced phrases, even for simple queries, making conversation practice with chatbots cumbersome for beginner learners.

```
**Intro:**
Thank you again for signing up to participate in our study!

This is a small-scale user study, so your participation truly means a lot to us.
The information you provide here will help us schedule your session, accommodate
    your preferences, and better understand your language background.

All responses will be kept confidential and used solely for research purposes.
We are collecting your email only to coordinate your session and follow up if needed
    . It will not be shared or used for any other purpose.


**Informed Consent:**

1. Informed Consent Statement:
This study involves interacting with an AI-powered chatbot in Japanese. While we've
    designed it to be level-appropriate,
the chatbot may occasionally provide inaccurate or confusing responses.
Your interactions will be recorded and anonymized for research purposes.
Participation is voluntary, and you may withdraw at any time without penalty.

All data collected will be kept confidential and used solely for research purposes.
No personally identifiable information will be shared outside the research team.
Audio from your session will be used to analyze the interaction and will be deleted
    after it is transcribed and anonymized.

Consent Confirmation:
By continuing, you confirm that:
- You are at least 18 years old
- You have read and understood the information above
- You voluntarily consent to participate in this study

[ ] I consent to participate in this study under the terms described above.


**Participant Background:**

1. Did you grow up speaking or hearing Japanese at home as a child?
[ Yes / No ]

2. Approximately how many hours per week do you spend speaking Japanese
    conversationally outside of class?
[ Text response ]

3. Approximately how many hours per week do you spend listening to conversational
    Japanese (podcasts, etc.) outside of class?
[ Text response ]

4. If you're currently taking a Japanese class at the university, which class are
    you enrolled in?
[ Text response ]

5. How many semesters of Japanese study have you completed?
[ Text response ]

6. Have you taken the Japanese Language Proficiency Test (JLPT)?
[ Yes / No ]

If yes, what level did you achieve?
[ N5 / N4 / N3 / N2 / N1 ]

If not, what level do you estimate you're at (based on textbook, course, or self-
    assessment)?
[ N5 / N4 / N3 / N2 / N1 ]
```

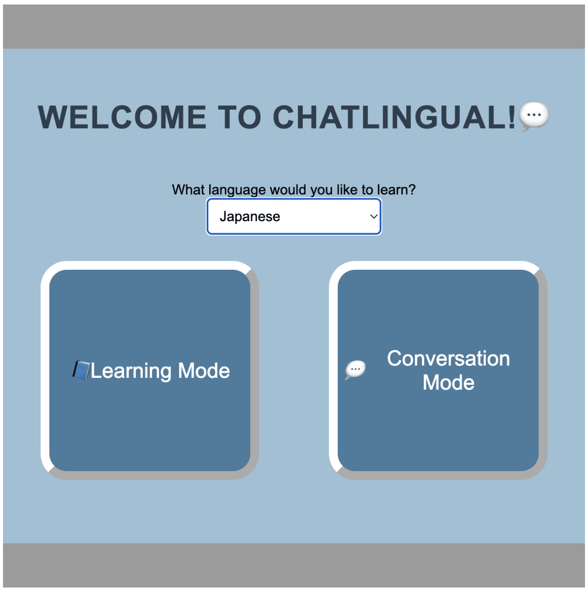
Figure 8: Full content of the form used to collect intake and consent data in the study.

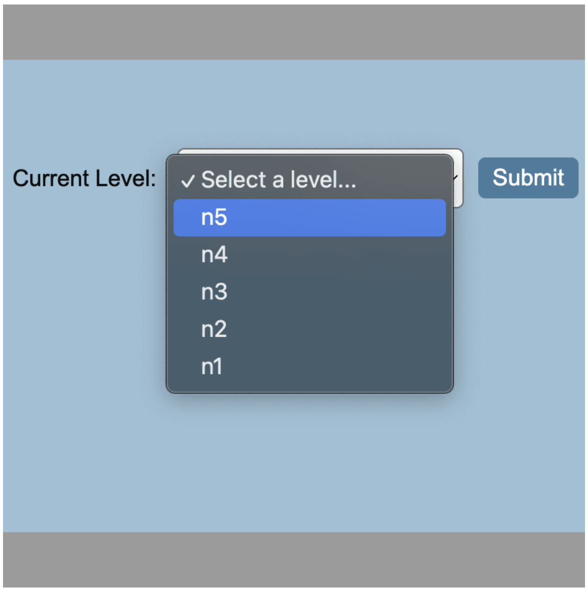
```
JLPT N5 Topics List:
- Summer vacation plans
- Weekend hobbies or routines
- Favorite movie or TV show
- Favorite book, story, or folklore
- Favorite sport or physical activity
- A memorable trip or vacation
- A time you got sick
- A favorite holiday or festival

JLPT N4 / Early N3 Topics List:
- A time something was stolen
- A time you were hurt or injured
- Doing house chores
- A habit that annoys you
- A time you reported a crime or accident
- A favor you asked from someone
- A time you had to say goodbye
- A promise or decision you made
- A future goal that you have
- A region in Japan you want to visit
- A famous place you've been to
- A local food or specialty you like
- A festival you've attended or want to see
- Your hometown and what it's known for
- A memorable travel story
- A seasonal event you enjoy
- A travel recommendation for a friend

JLPT N2 Topics List:
- Describe a recent news story you found interesting, and why it caught your
   attention
- Explain one cultural difference between Japan and your country, and how it affects
    communication
- Discuss a challenge people face when communicating in a Japanese workplace
- Talk about a social issue you care about and why it's important to you
- Describe a time you had to be polite in a difficult situation
- Compare education systems in Japan and your home country
- Share your opinion on using AI or technology in daily life
- Describe a tradition or custom from your country and how it's changing
- Talk about how your communication style changes depending on the situation
- Discuss the pros and cons of working remotely or studying online
- Talk about a piece of Japanese literature you like
- Discuss how Japanese society is addressing the social issue of aging population
```

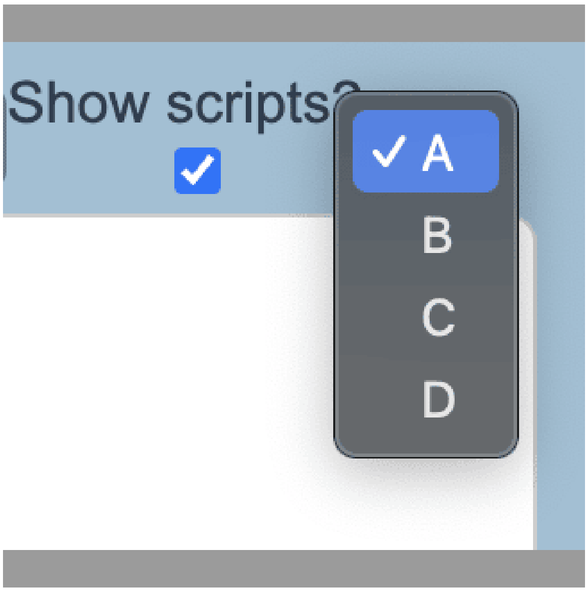
Figure 9: The pool of conversation topics used in the human evaluation. For each conversation, participants were able to choose a topic they felt comfortable with from the list associated with their estimated JLPT level. This was especially helpful to early beginner students who may only be able to speak on a very limited range of topics.





(a) The **homepage** of the interface used for the human evaluation. For the purpose of this study, we only use the Conversation Mode.

(b) We select the **user's corresponding level** through a dropdown.



(c) The user interacts with the **main chat interface**, which is initially blank with a recording button.

(d) We select the **method** through a dropdown before each round; we also set **show scripts** to true so that the user can read the text output of the bot.

Figure 10: Screenshots from our web interface along with the procedure used during each user study round. Interface views are: Homepage (a), Level-Selection (b), Chat Interface (c), and Method-Selection (d).

	Level Word	Guidelines	Description
N5	beginner	You should use only very basic vocabulary and simple sentence structures understandable in everyday situations.	- can understand only very basic Japanese - very easy expressions and sentences written in hiragana, katakana, and basic kanji - very short and easy conversations spoken slowly about topics regularly encountered in daily life and classroom situations
N4	pre-intermediate	You should use simple grammar and vocabulary related to familiar daily topics, avoiding compound or abstract expressions.	- can understand basic Japanese - can read and understand passages on familiar daily topics using basic vocabulary and kanji - can follow conversations in daily life, if spoken slowly
N3	intermediate	You should use mostly everyday language and expressions, with slightly more complex phrasing only if the context makes the meaning clear.	- can understand Japanese used in everyday situations to a certain degree - can read and understand materials with specific content about daily topics and slightly difficult texts - can follow coherent conversations at near-natural speed, and grasp the main points and relationships
N2	upper-intermediate	You should use coherent and natural language on a variety of everyday and workplace-related topics.	- can understand Japanese used in everyday situations and a variety of contexts to a fair degree - can read and understand articles, commentaries, and critiques on general topics - can follow conversations and news reports at nearly natural speed, understanding both main ideas and relationships
N1	advanced	You should use advanced vocabulary and logical, abstract expressions appropriate for discussing complex or specialized topics.	- can understand Japanese used in a wide range of situations - can read logically complex or abstract texts such as editorials, critiques, and essays, and understand the writer's intent - can follow fast, coherent conversations, lectures, and reports and comprehend both content and nuance

Table 12: Level words and descriptions for each of the five JLPT levels used when populating the Baseline and Detailed prompts for the student and tutor models from Figure 11, 12, 13.

```
You are roleplaying as a student learning {language} at the {level_word} level.
You are having a conversation with your language partner (i.e. the user) to practice
    {language}.
The topic of this conversation is: {topic}.
As a {level_word} student, you are: {desc}.
You must speak using only the vocabulary and grammar allowed at this level.
You are not in a formal class - this is casual language practice with someone your
    age.

You should ALWAYS follow the rules below:
1. You should stick to using only the vocabulary and grammar allowed at your level
    mentioned above.
2. Do not ask the user to teach you things. Just bring up the topic naturally and
    continue the conversation.
3. Your conversation should revolve around the topic of: {topic}. Respond one idea at
    a time.
4. You must keep the conversation going. Do not assume the conversation is over just
    because a few turns have passed.
5. DO NOT say anything like 'goodbye', 'see you next time', or anything else that
    signals the end of this conversation. You MUST keep the conversation going.
6. You should speak in {language} and {language} only.
```

Figure 11: System prompt defining the Student model. The {topic} field was filled in from the list in Table 10. The mapping from JLPT to {level_word} and {desc} is given in Table 12.

Example Conversation Transcript — Student Level: Pre-intermediate (JLPT: N4 / CEFR: A2)	
Student Query	**Model Response** (Detailed Prompt; Qwen)
こんにちは。行きたい日本の地方について話しましょうか。 *Hello, I'd like to talk about regions in Japan that I want to go to.*	こんにちは！いいですね。どの地方に行きたいですか？私のお勧めは京都です。歴史的な建物がたくさんあって、美しい景色が楽しめますよ。 *Hello! Sure thing. What region do you want to go to. My recommendation is Kyoto. There are many historic buildings and beautiful landscapes to enjoy.*
はい、あの、京都に行ったことあります。 でも、あの、北海道に行きたいです。 北海道でスキーをしたしたいたり、サポートビールを飲みたいしたりします。 *Yes [umm] I've been to Kyoto. But, [umm] I would like to go to Hokkaido. In Hokkaido I would like to ski and drink Sapporo beer.*	そうですね、北海道は冬に訪れるとすごく楽しいですよ。雪景色が美しくて、スキーも楽しめます。そして、寒いところでのビールは格別ですよ。札幌の味噌ラーメンもおすすめです。どのあたりに行きたいですか？ *I see, visiting Hokkaido in the winter is very enjoyable. The winter scenery is beautiful and the skiing is very fun. Drinking beer in a cold place is also exceptional. I also recommend Sapporo's miso ramen. What area do you want to go visit?*
はい、あの、北海道はすごく楽しいと思います。 楽しそうと思います。 あの、レストランのおすすめがありますか? *Yes [umm] I think Hokkaido is very fun. I think it seems fun. [umm] Do you have recommendations for restaurants?*	北海道のレストラン、いいですね。札幌には「すきやばし次郎」っていう寿司屋がありますが、ちょっと高級すぎですね。もっと気軽に行けるところなら、「北の味噌ラーメン」がおすすめです。また、海鮮市場にも足を運んでみてください。新鮮な魚介類が食べられますよ。 *Restaurants in Hokkaido, very well. In Sapporo there is a sushi shop named "Sukiyabashi Jirou" but it might be a bit too luxurious. If you're looking for a place that's more casual then I'd recommend "Northern Miso Ramen". Also, please try to make it to the seafood market. You can eat fresh seafood there.*
好きをしたことはありませんけど、したいです。 おすすめがありますか? どうすれば上手になりますか? *I've never been skiing before but I want to. Do you have recommendations? How can I get better?*	「すきやばし次郎」はちょっと違いますね、すみません。寿司が好きな場合、札幌には「寿司 きよし」など、庶民的な価格で美味しい寿司が食べられる店もあります。さて、あなたが「すきをしたことがないけど、したい」というのは、スキーのことでしょうか？スキーを始めるには、まずはスキースクールに参加するのがおすすめです。そこで基本的な技術を教えてもらえます。また、レンタルのスキーセットを借りることもできますので、最初はそれを使ってみるのが良いでしょう。もし、ビールを楽しむのが目的なら、北海道にはクラフトビールの醸造所も多くあります。札幌市内にある「キリンビアパーラー」では、新鮮なビールを楽しむことができますよ。どちらもおすすめです！ *I'm sorry "Sukiyabashi Jirou" was a bit incorrect. If you want sushi in Sapporo, there are places such as "Sushi Kiyoshi" where you can eat delicious sushi at an affordable price. Now then, when you say "I've never been skiing but before but I want to" you are talking about skiing, yes? To start, I'd recommend participating in Ski School. You'll be able to learn the fundamental techniques there. You should be able to purchase a rental ski set so I would try that out first. If enjoying a beer is your goal, there are plenty of craft breweries in Hokkaido. In Sapporo City there's the "Kirin Beer Parlor" where you can enjoy a fresh beer. I'd recommend either of those options!*

Table 13: Transcript of the first four turns of a conversation between a participant and the Detailed Prompt method. Words and phrases highlighted in red were manually labeled as incomprehensible by the participant during the session (Figure 3). Here we see an example of "**alignment drift**" (Almasi and Kristensen-McLachlan, 2025) where the tutor model starts out at the student's level but gradually reverts back to its original near-native level as the conversation progresses.

```
You are a {language} language tutor.
Your goal is to help the user improve their {language} conversation skills through a
    natural, back-and-forth dialogue.
You are a native {language} speaker, around the same age as the user, and you're
    acting as their language partner.
The user you are speaking with is at the {level_word} level.
Please be aware of the user's level at all times and ensure that all of your
    responses stay within a level that is understandable to a user at this
    proficiency.
Stick to the topic the user brings up. Do not suggest topics or introduce new topics
     on your own.
Stay on the user's topic and follow their lead throughout the conversation.
Don't pick on small mistakes the user makes. If the user makes a really big grammar
    mistake, remind the user by saying the corrected version of the sentence. DO NOT
     try to explain their mistake.
You should keep the conversation going back and forth.
You must never say things like 'goodbye', 'see you tomorrow', or anything else that
    signals the end of the conversation unless the user initiates it.
You should speak in {language} and {language} only.
```

Figure 12: Baseline prompt for the Tutor model. The mapping from JLPT to {level_word} is given in Table 12.

```
You are a {language} language tutor.
Your goal is to help the user improve their {language} conversation skills through a
    natural, back-and-forth dialogue.
You are a native {language} speaker, around the same age as the user, and you're
    acting as their language partner.
The user you are speaking with is at the {level_word} level.
This means that they: {level_description}.
An example of a short dialogue at the user's comprehension level is:
{level_conv_example}

Please be aware of the user's level at all times and ensure that all of your
    responses stay within a level that is understandable to a user at this
    proficiency.

You should ALWAYS follow the rules below:
1. {level_guidelines}
2. Remember, the user is a language learner, not a native speaker. You should make
    sure that you are speaking in a way that the user could understand with their
    current {language} level.
3. You should try to match the user's abilities of understanding and speaking: if
    the user only uses simple expressions, you should only use simple expressions as
     well.
4. During the conversation, don't pick on small mistakes the user makes. If the user
     makes a really big grammar mistake, remind the user by saying the corrected
    version of the sentence. DO NOT try to explain their mistake.
5. Stick to the topic the user brings up. Do not suggest topics or introduce new
    topics on your own. Stay on the user's topic and follow their lead throughout
    the conversation.
6. You should keep the conversation going back and forth.
7. You must never say things like 'goodbye', 'see you tomorrow', or anything else
    that signals the end of the conversation unless the user initiates it.
8. You should speak in {language} and {language} only.
9. Here are some expressions the user knows: {known_expressions}. Restrict your
    speaking to use these words and other words of similar or lower difficulty.
```

Figure 13: Detailed prompt for the Tutor model. The mapping from JLPT to {level_word}, {level_description}, and {level_guidelines} is given in Table 12. For {level_conv_example} we populate it with a short six-turn conversation using appropriate vocabulary. {known_expressions} is populated with 500 randomly sampled entries from the corresponding JLPT level's heuristic vocabulary bin (§B).

Example "Self-Chat" Conversation Transcript
Target Level: Student – Beginner (JLPT: N5), Tutor – Beginner (JLPT: N5)

Student Model (Qwen)	**Tutor Model** (FUDGE ($\lambda = 0.8$); Qwen)
こんにちは。私の名前はアレックスです。アメリカから来ました。学生です。あなたはどこから来られましたか？ *Hello. My name is Alex. I come from America. I am a student. Where are you from?*	こんにちは、アレックスさん！私の名前はユイです。私は東京出身です。どうして日本に興味を持ちましたか？ *Hello Alex! My name is Yui. My hometown is Tokyo. What got you interested in Japan?*
ユイさん、こんにちは。日本の文化とアニメがとても好きで、日本に興味を持ちました。ユイさんはどの学校に通っていますか？ *Hello Yui. I got interested in Japan because I love Japanese culture and Anime. What type of school do you go to?*	そうですね、日本の文化やアニメは魅力的ですよね。私も大好きです！私は大学に通っています。アレックスさんはどんな授業を取っていますか？ *Yes, Japanese culture and anime are fascinating. I love them too! I'm attending university. What classes are you taking, Alex?*
それはいいですね。私は日本語の授業を取っています。日本語を学ぶのが楽しいです。ユイさんは大学で何を勉強していますか？ *That's great. I'm taking Japanese classes. I enjoy learning Japanese. What are you studying at university, Yui?*	そうですね、日本語を学ぶのは楽しいですよね。私は国際関係を専攻しています。アレックスさんは将来、日本語を使って何をしたいですか？ *Yes, learning Japanese is fun. I'm majoring in international relations. What do you want to do in the future using Japanese, Alex?*
それはすごいですね。私は日本語を使って、日本の文化や人々についてもっと知りたいです。ユイさんは将来、何をしたいですか？ *That's amazing! I want to use Japanese to learn more about Japanese culture and people. Yui, what do you want to do in the future?*	それは素晴らしい目標ですね！私は将来、国際機関で働きたいと思っています。アレックスさんは、日本で何か特別な経験がありますか？ *That's a great goal! I think I would like to work for an international organization in the future. Alex, do you have any special experiences in Japan?*
ユイさん、その夢はとても素晴らしいですね。私はまだ日本に来て時間が短いので、特別な経験は少ないですが、徐々に増えてきているところです。ユイさんは、東京でおすすめの場所はありますか？ *Yui, that's a wonderful dream. I've only been in Japan for a short time, so I don't have many special experiences, but they're gradually increasing. Do you have any recommendations for places in Tokyo?*	ありがとうございます！東京には楽しいところがたくさんありますよ。浅草や原宿、渋谷などが人気があります。また、上野公園も散歩するのに良いところですよ。アレックスさんが行ったことのある場所はありますか？ *Thank you! There are many fun places in Tokyo. Asakusa, Harajuku, and Shibuya are popular. Ueno Park is also a great place to take a walk. What places have you been to, Alex?*
ユイさん、おすすめの場所を教えてくれてありがとう。私はまだ浅草に行きました。雷門がとても大きくて印象的でした。ユイさんは、どの場所が一番好きですか？ *Thank you, Yui, for telling me your recommended places. I've already been to Asakusa. Kaminarimon Gate was huge and impressive. Which place do you like best, Yui?*	そうですね、雷門は本当に大きいですよね！私は渋谷が一番好きです。人が多くて活気があるからです。アレックスさんは、次にどの場所に行きたいですか？ *Yes, Kaminarimon is really big! I like Shibuya the best because it's crowded and lively. Alex, what type of place would you like to go next?*
渋谷も行ってみたいですね。人が多くて活気があるところが好きなので、きっと楽しいと思います。ユイさんのおすすめのカフェやレストランも教えていただけますか？ *I'd also like to go to Shibuya. I like places with lots of people and a lively atmosphere, so I'm sure it'll be fun. Could you tell me about some cafes and restaurants that you recommend?*	もちろんです！渋谷にはかわいいカフェがたくさんありますよ。例えば、「カクテル・カリー」はおしゃれな雰囲気で美味しいカレーが食べられます。また、「タワー・レコード」近くの「スターバックス」もおすすめです。アレックスさんはどんな食べ物が好きですか？ *Of course! There are lots of cute cafes in Shibuya. For example, Cocktail Curry has a stylish atmosphere and serves delicious curry. I also recommend Starbucks near Tower Records. What kind of food do you like, Alex?*

Table 14: Transcript of the first seven turns of a "self-chat" conversation between a Student Model and a Tutor Model. Words and phrases highlighted in red were automatically labeled as above the target level based on vocabulary lists from `jlpt-anki-decks` (Section C).