Dango: A Strictly L1-Only Large Language Model for Studying Second Language Acquisition

Shiho Matta[1] **Yin Jou Huang**[1] **Fei Cheng**[1] **Takashi Kodama**[2]
Hirokazu Kiyomaru[2] **Yugo Murawaki**[1]
[1]Kyoto University [2]NII-LLMC
{matta,huang,feicheng,murawaki}@nlp.ist.i.kyoto-u.ac.jp
{tkodama,kiyomaru}@nii.ac.jp

Abstract

We introduce **Dango**, a 1.8B-parameter large language model designed for controlled studies of L1→L2 (Japanese→English) transfer in second language acquisition (SLA). While previous studies have explored SLA in language models, they have predominantly relied on smaller or non-decoder models, limiting their ability to generate open-ended text and reducing their suitability as practical L2 simulators. We identify a key challenge when scaling models to this size: L2 contamination within the "monolingual" pretraining corpus used for L1 acquisition. To address this, we propose a filtering method to reduce premature exposure to English while preserving realistic, minimal exposure. We then fine-tune the model on LLM-generated L2-learning lessons to simulate the L2 acquisition process. Our evaluations show that Dango develops human-like L2 production patterns, reliably matching Japanese learners of English more closely than the multilingual baseline across our distributional metrics and the unfiltered baseline on our usage frequency metric. We release the model, data, and code to facilitate reproducible computational SLA research and learner-facing applications.[1]

1 Introduction

Second language acquisition (SLA) refers to the learning of a second language (L2) after a first language (L1) has already been acquired natively. In humans, the influence of L1 can facilitate L2 learning in some cases and interfere with it in others, a phenomenon widely known as the language transfer effect (Lado, 1957; Odlin, 1989). This may result in overuse, avoidance, and misuse patterns of certain linguistic features in a speaker's L2 production (Jarvis and Pavlenko, 2008). For example, because Japanese lacks a grammatical article system, Japanese-speaking learners of English often show difficulty with English article use (Murakami and Alexopoulou, 2015). This can lead to article-omission errors such as "I went to library" instead of "I went to the library."

Figure 1: English contamination examples: severe case of translation pairs versus minor, typical English exposure for Japanese speakers.

As transformer-based language models (LMs) have recently shown strong performance on many linguistic tasks (Lai et al., 2023), many studies have investigated whether they also exhibit such human-like transfer effects (Oba et al., 2023; Aoyama and Schneider, 2024; Yadavalli et al., 2023; Constantinescu et al., 2025). These studies train LMs from scratch first on L1 and then on L2, thereby mimicking the order in which humans learn languages. Transfer effects are then evaluated, for example, by testing the model's grammatical judgments on the L2 and comparing them with human tendencies (Yadavalli et al., 2023).

Although these studies provide valuable insights into transfer effects in LMs, their use as practical simulators of L2 speakers remains limited. This is largely due to limitations of architecture and scale. Prior work has used either encoder-only models (Oba et al., 2023; Yadavalli et al., 2023; Constantinescu et al., 2025) or small decoder-only models, such as the 137M-parameter GPT-2 in Constantinescu et al. (2025). As a result, these models are less suitable for robust open-ended or interactive generation, and for potential applications such as L2 learner support (Lyu et al., 2024) and teacher

[1]https://github.com/mattashiho233/dango

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 59856–59876
October 24-29, 2026 ©2026 Association for Computational Linguistics

training with virtual L2 speakers (Vasmatzoglou and Ní Chiaráin, 2020).

To equip future SLA research with a more modern and capable model, and to enable broader practical applications, we release **Dango**, a Llama-2-like decoder-only transformer trained in an L1→L2 manner at a much larger scale. Dango has 1.8B parameters and is pretrained on far more data than prior work, with up to 100B L1-stage tokens. In this study, we focus on Japanese as the L1 and English as the L2.

To train an LLM at this size, web-crawled text is the only viable data source (Hoffmann et al., 2022). However, the Japanese web corpus we used for L1 acquisition contained substantial accidental English exposure. We call this problem **L2 contamination**. Although native Japanese speakers naturally see English through loanwords, proper nouns, and code-switching (Hyde, 2002; Stanlaw, 2004), systematic acquisition of English is relatively rare outside classroom-like educational settings (Barrs, 2020). It is therefore natural to reshape the corpus into a state closer to a typical Japanese speaker's environment: removing structured English content, such as in the worst case, translation pairs, while retaining minimal English such as brand names and simple phrases (Figure 1). We designed a filtering pipeline that removes excessive English exposure, reducing the corpus size by 30% (Figure 2).

In the L2 acquisition stage, we fine-tune Dango on LLM-synthesized textbook-like lessons. The lessons follow a gradual difficulty progression, as they are based on an English seed word drawn from a word list with difficulty labels. The lessons are anchored in both Japanese and English, containing (i) a Japanese explanation of the seed word, (ii) an example English sentence using the word, paired with a Japanese translation, and (iii) a brief Japanese commentary highlighting relevant grammar or usage. In this way, Dango is exposed to English while keeping instruction and metalinguistic reasoning in its L1, which mirrors common textbook learning practice (Nakayama, 2022).

Our experiments show that Dango acquires English generation abilities, including the capacity to produce long English sentences absent from its strictly filtered pretraining data. Evaluations further show that Dango is reliably closer to human Japanese-English learners than the multilingual baseline on both usage frequency and error rate, and closer than the unfiltered baseline on usage frequency, while performing comparably to prompted GPT-5.5. These results provide evidence of successful language transfer. We release Dango's code and checkpoints to support controlled SLA research and practical applications.

2 Related Work

2.1 Second Language Acquisition (SLA) with Language Models (LMs)

Modeling SLA by Training in an L1→L2 Setup. Previous studies have trained LMs on L1 and L2 sequentially to model human L2 acquisition. Oba et al. (2023) trained 18M-parameter encoder-only models and analyzed how different L1 pretraining languages shape the trajectory and ultimate success of L2 grammatical generalization. Yadavalli et al. (2023) also trained 110M-parameter encoder-only models, but aimed to isolate positive and negative cross-linguistic transfer. Aoyama and Schneider (2024) trained 11M-parameter (our estimation) GPT-2 models as L2 English speakers, investigating transfer effects and showing that LM surprisal helps predict L2 English readers' word-by-word reading times. Constantinescu et al. (2025) trained 125M-parameter encoder-only and 137M-parameter decoder-only models, but their focus was on recreating the critical period effect observed in humans rather than directly examining transfer effects.

Training Data. Training data is central to modeling SLA. Some work has attempted to create a more human-like setting by using data derived from human speech. For example, Yadavalli et al. (2023) used child-directed speech, and Constantinescu et al. (2025) partially used spoken-language data. However, these data sources are limited in size, and Constantinescu et al. (2025) had to include web text to obtain sufficient data and cross-linguistic coverage. On the other hand, Oba et al. (2023) and Aoyama and Schneider (2024) relied on CC-100 (Conneau et al., 2020; Wenzek et al., 2020) web-crawled data for L1 training.

Evaluating Transfer Effect. Oba et al. (2023), Aoyama and Schneider (2024), Yadavalli et al. (2023), and Constantinescu et al. (2025) used BLiMP (Warstadt et al., 2020) to measure the L2 (English) grammatical generalization. Other probes include perplexity (Aoyama and Schneider, 2024; Constantinescu et al., 2025), model surprisal (Aoyama and Schneider, 2024) and knowledge-based multiple-choice tasks (Constantinescu et al.,

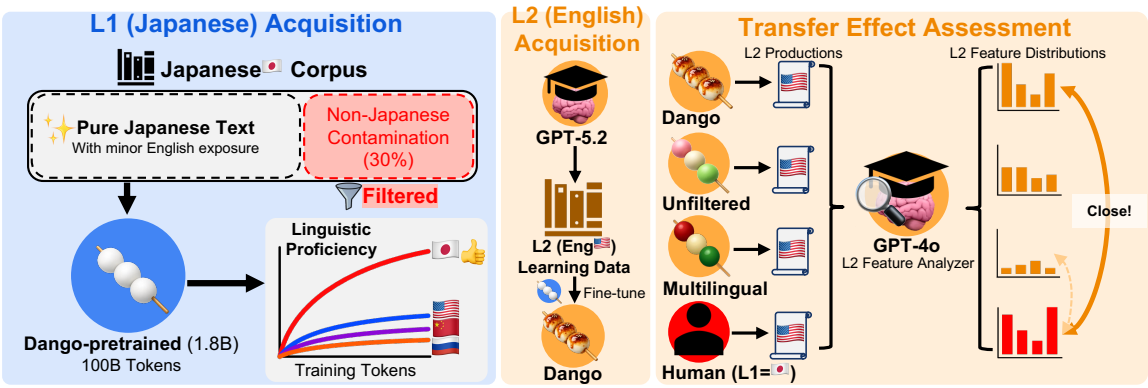


Figure 2: An overview of our proposed methodology.

2025).

However, because these probes rely entirely on the predefined inputs within existing benchmarks, they do not truly test a model's generation ability. To address this limitation, we evaluate the transfer effect in Dango's English text production using the automatic LLM-as-a-judge framework introduced by Gao et al. (2025). We directly annotate linguistic features in the model's output to compare their distribution with authentic human L2 productions.

2.2 Controlling Knowledge Exposure during Pretraining

LLMs with Knowledge Cutoff. Recent efforts highlight the necessity of strictly constrained pretraining environments to isolate specific phenomena from the massive data contamination inherent in standard LLMs. For instance, TimeCapsule LLM is trained exclusively on texts from constrained historical periods to accurately emulate the vocabulary and worldview of that era without the intrusion of modern concepts.[2] Similarly, History LLMs are trained from scratch using strict temporal knowledge cutoffs to completely prevent the leakage of future knowledge.[3] These projects demonstrate that true behavioral simulation requires a verifiable, uncontaminated training baseline.

L2 Contamination in Web Corpora. L2 contamination matters in controlled studies of second-language acquisition and transfer effects. Because language identification is imperfect, corpora intended to be non-English often still contain English data (Caswell et al., 2020). In controlled experiments, such accidental exposure can be problematic because it may introduce unintended cross-linguistic transfer (Blevins and Zettlemoyer, 2022). This issue may also apply to prior work that used web corpora for training (Oba et al., 2023; Aoyama and Schneider, 2024; Constantinescu et al., 2025), although any resulting failures may have been too subtle to detect at their scale. To our knowledge, our work is the first to directly address the L2 contamination problem in computational SLA.

3 Methodology

The training of Dango consists of two stages (Figure 2). First, we model L1 (Japanese) acquisition by pretraining a decoder-only language model on a strictly filtered Japanese corpus (§ 3.1), which yields the **Dango-pretrained** model. Second, we model L2 (English) acquisition by fine-tuning the pretrained model on LLM-generated L2 learning data, yielding **Dango** (§ 3.2).

3.1 L1 Acquisition

We conducted filtering to remove non-Japanese exposures in a Japanese corpus and trained a decoder-only transformer model on the resulting data.

3.1.1 Corpus for L1 Acquisition and Filtering

To bootstrap our study from an established baseline, we adopted llm-jp-corpus-v3, the training corpus released by the llm-jp project.[4] The full corpus is multilingual, containing Japanese, English, Chinese, Korean, and code data. In this work, we used only the Japanese portion, which is mainly composed of Japanese Common Crawl data[5] and PDF data from the National Diet Library's Web Archiving Project (WARP).[6] Nevertheless, we observed that this subset still included a non-negligible amount of English text.

[2] https://github.com/haykgrigo3/TimeCapsuleLLM
[3] https://github.com/DGoettlich/history-llms
[4] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3/-/tree/main
[5] https://commoncrawl.org.
[6] National Diet Library (Japan). NDL Web Archiving Project (WARP), https://warp.ndl.go.jp/.

We built a filtering pipeline to limit exposure to English grammar and long, structured English sentences that could confound L1→L2 transfer (Barrs, 2020; Blevins and Zettlemoyer, 2022), while retaining benign English signals common in Japanese text, such as everyday words, phrases, and proper nouns (Hyde, 2002; Stanlaw, 2004).

The pipeline combines a whitelist filter with blocklist filters applied at both the document and line levels. The whitelist filter defines the character inventory allowed for Japanese-centric text, including Japanese scripts, common CJK characters, punctuation, symbols, and other frequent marks. Documents with more than 1‰ out-of-inventory characters are discarded, which removes most text in non-CJK and non-Latin scripts, such as Russian and Arabic. Next, blocklist filters remove lines separated by "\n" that contain excessive English or Chinese characters. For example, the English filter removes a line if it contains more than four consecutive English words, more than 20 Latin letters, or a Japanese-to-English character ratio below 40%. If more than 5% of a document's lines are removed, the entire document is discarded to avoid damaged context. Further implementation details are provided in Appendix A.

3.1.2 Pretraining an L1-Only LLM

We pretrained a 1.8B-parameter Llama 2–style decoder-only Transformer from scratch on the corpus described in § 3.1.1. To our knowledge, this is the largest model trained from scratch for studying second language acquisition. Pretraining details are provided in Appendix C.

We adopted the pretrained multilingual tokenizer from the released llm-jp-3 model, which uses a vocabulary of 99,584 tokens covering both Japanese and English. We acknowledge that this gives the model access to L2 subword units from initialization. We leave experiments with more human-like tokenization assumptions to future work.[7]

In this work, "strictly L1-only" refers to explicitly controlling sustained L2 exposure through corpus filtering, rather than excluding every English token: the tokenizer contains English subwords, and the filtered corpus intentionally retains isolated English words, proper nouns, and short phrases.

3.1.3 Linguistic Proficiency Assessments

We monitor the model's linguistic proficiency during L1 pretraining to assess whether it effectively acquires Japanese and whether unwanted language contamination occurs after filtering.

Assessing Japanese Proficiency. We assess Japanese proficiency along two dimensions: grammatical knowledge and general Japanese-language ability. For grammatical knowledge, we use JBLiMP (Someya and Oseki, 2023), a Japanese benchmark in the style of BLiMP (Warstadt et al., 2020). Models are evaluated on minimal pairs that differ in a targeted grammatical property; an item is correct if the model assigns higher probability to the grammatical sentence. Performance is reported as accuracy, with chance at 50%.

For general Japanese-language ability, we use llm-jp-eval-v1.4.1 ,[8] a few-shot benchmark (Brown et al., 2020) covering diverse tasks such as natural language inference and multiple-choice question answering. Programming and mathematical tasks are excluded.

Assessing English Leakage. We use the English section of MultiBLiMP (Jumelet et al., 2025), another BLiMP style benchmark.

Assessing Other Leakage. We assess Russian and Chinese proficiency using the Russian part in MultiBLiMP and ZhoBLiMP (Liu et al., 2025).

3.2 L2 Acquisition

To introduce Dango to English, we synthesized English learning data using an off-the-shelf multilingual LLM and fine-tuned the model.

3.2.1 Synthesizing L2-Learning Data with LLM

We generate parallel, lesson-style data for English L2 learning using a strong multilingual LLM. Each example is centered on a seed word from the CEFR-J vocabulary list (Tono, 2019), a CEFR-based (Council of Europe, 2001) inventory tailored to Japanese learners. Since the original CEFR-J annotations cover only A1–B2, we use an externally annotated extension to include words up to C2. We refer to the combined inventory as CEFR-J throughout the paper.[9]

As shown in Figure 13, we prompt the LLM to generate: (i) a detailed explanation of the seed word, (ii) an example sentence containing the word

[7] https://huggingface.co/llm-jp/llm-jp-3-1.8b

[8] https://github.com/llm-jp/scripts/tree/main/evaluation/installers/llm-jp-eval-v1.4.1

[9] https://github.com/openlanguageprofiles/olp-en-cefrj/tree/master , C1/C2 labels are from the Octanove Vocabulary Profile (CC BY-SA 4.0).

and its Japanese translation, and (iii) an analysis of salient usage patterns in the example. Since the prompt controls sentence difficulty, average sentence length and complexity increase with the CEFR-J level of the seed word (Figure 16). Both the lexical explanation and usage analysis are written in Japanese, reflecting textbook-style English learning for Japanese speakers and encouraging Japanese-to-English transfer.

The above setup yields the **English–Japanese** (Eng–Jpn) setting, where the model generates an English sentence followed by its Japanese translation. We also create a **Japanese–English** (Jpn–Eng) variant by simply swapping the English and Japanese sentences in the data.

3.2.2 Fine-tuning on L2 Learning Data

We simulate L2 acquisition by fine-tuning models on L2 learning data. Each training example follows a prompt format in which the user asks for an explanation of a word and the assistant provides the corresponding L2 learning response (Figure 14). By combining data from different CEFR levels, we train models that mimic L2 learners at specific proficiency stages (see § 5.1).

3.2.3 Assessing L2 Acquisition Progress

We evaluate L2 acquisition progress with BLEU scores on Eng–Jpn and Jpn–Eng translation. We use the translation component of a held-out subset of the L2 learning data for in-distribution testing, and Jpn–Eng parallel sentences from Tatoeba for out-of-distribution evaluation.[10]

3.2.4 L1–L2 Transfer Assessment

To directly quantify L1–L2 transfer effects, we adopt an LLM-as-a-judge evaluation protocol following Gao et al. (2025) to analyze the patterns in the English productions and compare that to human performance. We prompt a judge LLM to extract instances of seven English linguistic feature categories (excluding *Speech Acts*) from model outputs and to assess their contextual correctness: *Number Agreement* (local number concord inside noun phrases), *Tense Agreement* (tense choices consistent with temporal cues), *Subject–Verb Agreement* (verb inflection matching the subject), *Modal Verbs and Expressions*, *Quantifiers and Numerals*, *Noun–Verb Collocations* (conventional verb–noun pairings supporting fluent lexical selection), and *Reference Word* (referential devices such as pronouns/demonstratives that maintain discourse continuity).

As shown in Figure 18, the judge identifies each instance of a target feature in a translation output and returns (i) the exact text span and (ii) a binary label indicating whether the usage is correct in context. From these analysis outputs, we derive for each category (a) a **Usage Frequency (UF)** distribution (how often the feature appears) and (b) an **Error Rate (ER)** distribution (how often detected instances are incorrect).

Finally, we examine whether Dango produces English texts that share UF and ER distributions similarly to a human English learner. We apply the same llm-as-a-judge method to extract feature distributions from the ICNALE Written Essays (Ishikawa, 2013, 2023) corpus. This corpus contains topic-centric essays written by English learners of various L1 backgrounds, including Japanese. We use Jensen–Shannon divergence (JSD) to capture distributional shift in UF, and Bray-Curtis Dissimilarity (BCD) in addition to JSD in ER to better reflect raw error count shifts (Appendix J.1). We also provide an extended analysis comparing model L2 production against ICNALE Spoken Dialogues (Ishikawa, 2019) in Appendix J.2.

4 Experiments and Results for L1 (Japanese) Acquisition

In this section, we describe the experimental setup and evaluation results for pretraining Dango-pretrained, the L1-only LLM. For brevity, we may omit the "-pretrained" suffix when referring to the pretrained models in this section.

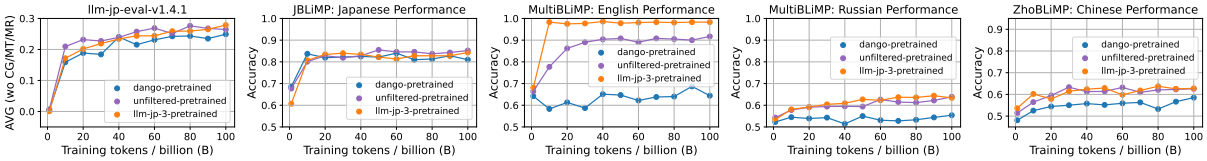
4.1 Experimental Settings

We detail the data preparation and training method for Dango-pretrained and the baselines.

Training-based Baselines. For training-based baselines, we trained an **unfiltered** model and used the **llm-jp-3** models as multilingual baselines. The unfiltered model was trained on the Japanese portion of llm-jp-corpus-v3 without the filtering described in § 3.1.1. We obtained the llm-jp-3-pretrained checkpoints by contacting the llm-jp team directly. These models have the same size, architecture, and training method as Dango; they differ only in the training data used.

L1 Corpus Filtering and Statistics. Detailed filtering parameters are provided in Appendix A.

[10] https://tatoeba.org . Created by volunteer contributors from the Tatoeba community. Text data licensed under CC BY 2.0 FR. Downloaded 2025-12-05.



(a) Japanese (llm-jp-eval-v1.4.1) (b) Japanese (JBLiMP) (c) English (MultiBLiMP) (d) Russian (MultiBLiMP) (e) Chinese (ZhoBLiMP)

Figure 3: Linguistic proficiency assessments across the L1-pretraining stage.

Filtering removed 30% of documents (Table 11); the remaining documents were further stripped of English-heavy lines through line-level filtering. Figures 8 and 9 show document- and phrase-level contamination removed by the English filter. However, short English phrases, such as common daily words and proper nouns, may remain (Figure 10). Thus, Dango can still learn weak word- or phrase-level English associations, but not the ability to generate long, coherent English sentences after L1 pretraining, as shown in Appendix E.

L1 Pretraining Details. Dango-pretrained was trained on 100B tokens from the filtered corpus. Compute-optimal scaling suggests that a 1.8B-parameter model is optimally trained on roughly 35B tokens (Hoffmann et al., 2022); however, we trained for about 3× more tokens because its Japanese generalization continued improving beyond 35B tokens (Figure 3a). Additional pretraining details are provided in Appendix C.

4.2 Linguistic Proficiency Assessments

We track the linguistic proficiency of the pretrained models over a 10B-token training interval.[11]

Japanese Proficiency. On llm-jp-eval (Figure 3a), Dango performs slightly below llm-jp-3 as training approaches 100B tokens (-0.029). A category breakdown is provided in Appendix D.

On JBLiMP (Figure 3b), Dango outperforms both baselines in early stage (1B and 10B tokens), likely because it is trained on nearly pure Japanese data. As training continues, the gap closes and all models reach comparable performance. Overall, Dango attains Japanese ability roughly on par with the unfiltered and llm-jp-3 models.

English Proficiency. On the English portion of MultiBLiMP (Figure 3c), llm-jp-3 saturates after only 10B tokens, likely because roughly half of its training tokens are English. In contrast, Dango remains near 0.65 even after 100B tokens. The unfiltered model indicates that naive training on a "Japanese" corpus can still yield strong English proficiency, reaching 0.917.

These results suggest that filtering prevents Dango's English ability from scaling with additional pretraining. The case study in Appendix E further shows that Dango's English remains largely n-gram-like, likely because the filtering rule removes sequences of more than four consecutive English words.

Other L2 Leakage. As an additional check for L2 leakage, we evaluate Russian and Chinese (Figures 3d and 3e) to assess the effectiveness of the whitelist and Chinese filter. All three models perform poorly, but Dango shows the lowest proficiency in both languages.

5 Experiments and Results for L2 (English) Acquisition

5.1 Experimental Settings

Synthesizing L2 Learning Data. We used GPT-5.2 (`gpt-5.2-2025-12-11`) to synthesize 9.9K textbook-style L2-learning examples from the CEFR-J word list, generating one example per target word. The dataset comprises 8,938 training, 494 validation, and 501 test examples. See Appendices F.1 and F.3 for prompt design and additional synthesis statistics.

Before fine-tuning on the L2 learning data, we first trained the model on Ichikara, a Japanese instruction-following dataset.[12] This step was intended to strengthen instruction-following behavior in the model's native language and to enable the case studies in § 4.2 (Appendix E).

Fine-tuning on L2 Learning Data. We trained separate models for six CEFR-J proficiency levels under two direction settings: English–Japanese (Eng–Jpn) and Japanese–English (Jpn–Eng). For a given target level, we used cumulative data from

[11] In Figure 3, the leftmost points on the plot correspond to models trained on 1B tokens.

[12] Visit this link for the Ichikara dataset (CC-BY-NC-SA).

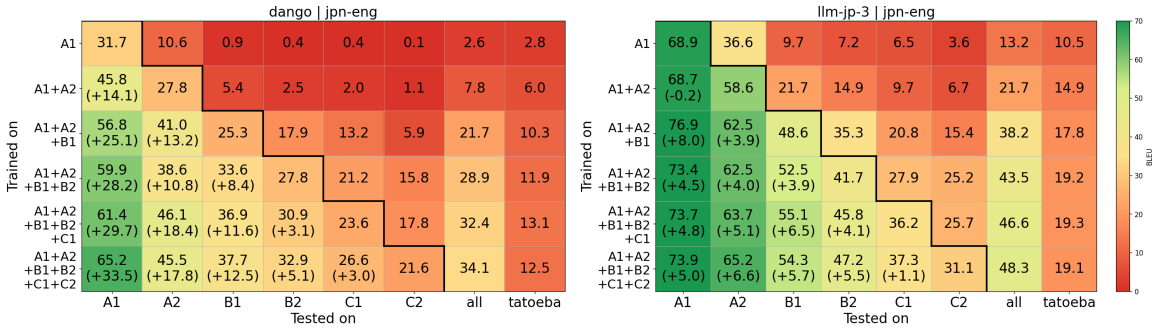


Figure 4: Dango and llm-jp-3 Japanese-to-English translation performance when trained and evaluated on different levels of data. The stair-shaped black line separates seen levels (lower left) from unseen levels (upper right).

all levels up to and including that level (e.g., B2 uses A1+A2+B1+B2). See Appendix F.2 and G for additional fine-tuning details.

5.2 L2 Proficiency Development Assessment

Figure 4 shows Jpn–Eng translation performance across difficulty levels for Dango and llm-jp-3 trained on different mixtures of difficulties. Results for the Eng–Jpn direction are shown in Appendix H. For Tatoeba, we randomly sampled 1,000 sentence pairs from the downloaded dataset. We provide a human analysis of whether the BLEU score captures meaningful L2 development in Appendix I.

5.2.1 Japanese–English Translation

General observations. When trained and evaluated on the same difficulty level, for example A1, Dango produces understandable English translations and reaches a BLEU score of 31.7. At this level, llm-jp-3 outperforms Dango with a +37.2 BLEU advantage.

Performance improves as the training and evaluation settings expand. When trained on all levels from A1 to C2 and evaluated on C2, Dango reaches 21.6 BLEU. Qualitatively, it often preserves the core meaning in long sentences, but with frequent lexical and grammatical errors (see Appendix I.1). Overall, llm-jp-3 consistently outperforms Dango, likely due to its multilingual pretraining.

Both models can generalize to unseen levels and Tatoeba data to some extent. For example, when trained on A1–B2 data, Dango generalizes to C1 (21.2) and C2 (15.8). Being multilingual, llm-jp-3 generalizes better, achieving 27.9 and 25.2.

Training on more difficult data improves performance at easier levels. Training with more difficult data consistently enhances performance on easier test sets. For instance, when evaluated on B1, Dango sees a +12.5 BLEU gain when training expands from A1–B1 to A1–C2, whereas llm-jp-3 gains only +5.7 BLEU. This demonstrates that Dango learns effectively from the L2-learning data.

Model	JSD of UF	JSD of ER	BCD of ER
Training-based methods			
Dango	0.0104 🥈3	0.0110 🥇1	0.2467 🥈2
Unfiltered	0.0176	0.0304 🥈3	0.2502 🥈3
llm-jp-3	0.0249	0.0487	0.6982
Prompting-based methods			
llm-jp-3.1-instruct	0.0346	0.0658	0.8523
GPT-4o	0.0028 🥇1	0.0374	0.5295
GPT-5.5	0.0084 🥈2	0.0191 🥈2	0.1925 🥇1

Table 1: LLM output linguistic feature distribution divergences to human Written Essay produced by Japanese learners. Lower is more similar.

Human L1	JSD of UF	JSD of ER	BCD of ER
Japanese	0.0104 🥇1	0.0110 🥈2	0.2467 🥈3
Korean	0.0111 🥈2	0.0094 🥇1	0.1078 🥇1
Mandarin	0.0140 🥈3	0.0161 🥈3	0.3509
Thai	0.0155	0.0221	0.1272 🥈2

Table 2: Linguistic feature divergences between Dango and humans with different L1 backgrounds.

5.3 Transfer Effect Assessment

Baselines. In addition to the training-based methods (Dango, unfiltered, and llm-jp-3), we adopt the prompting-based L2 speaker simulation method from Gao et al. (2025), where an LLM is instructed to generate L2 English text conditioned on prompts containing a specified L1 background and L1-dependent linguistic traits. We evaluate three LLMs in this setting: llm-jp-3.1-instruct,[13] GPT-4o (`gpt-4o-2024-11-20`),[14] and GPT-5.5

[13] We used the 1.8B model: https://huggingface.co/llm-jp/llm-jp-3.1-1.8b-instruct4

[14] https://openai.com/index/hello-gpt-4o/

(gpt-5.5-2026-04-23).[15] For GPT-4o and GPT-5.5, we follow the original prompt design, modifying it to output English translations (Figure 21 and 23); for llm-jp-3.1-instruct, we use a simplified role-playing and error-injection prompt (Figure 24) due to its smaller size and shorter context limit. We note that, because GPT-4o is also used as the annotation judge, its own baseline results may be affected by potential self-preference bias.

Both approaches offer distinct advantages. Prompting-based methods are highly accessible, utilizing off-the-shelf LLMs without the need for additional training. Conversely, training-based models allow us to trace the transfer effect directly back to the L1→L2 training process itself. This isolates the learning mechanism from the confounding effects of an LLM simply "role-playing" to mask its actual proficiency, providing a stronger foundation for future controlled studies and interpretability research.

As described in § 3.2.4, we first construct UF and ER distributions for human and model English productions. We then compute JSD and BCD for human–model pairs to identify which models best match human transfer patterns. Here, we report results on ICNALE Written Essays; Appendix J.2 presents the corresponding analysis on ICNALE Spoken Dialogues, providing a complementary robustness check. For trained baselines, we use models trained on A1–B2 data, since the ICNALE learners are at or below B2 level.

Focusing on JSD of UF. Table 1 shows that, among trained baselines, Dango is closer to Japanese L1 speakers than the unfiltered model or llm-jp-3 in terms of UF divergence. In other words, Dango has more human-like overuse and avoidance patterns over the seven feature categories. The distributions used to calculate these divergences are shown in Figure 5.

Compared with prompting baselines, GPT-4o and GPT-5.5 show UF distributions closer to humans, slightly outperforming Dango. In contrast, llm-jp-3.1-instruct remains the most distant, likely due to its small size and limited exposure to such role-play prompts during training. It generally failed to output systematic English grammar errors, even when prompted to do so.

Table 2 shows Dango's UF divergence from human English learners across L1 backgrounds. The

[15] https://openai.com/index/introducing-gpt-5-5/

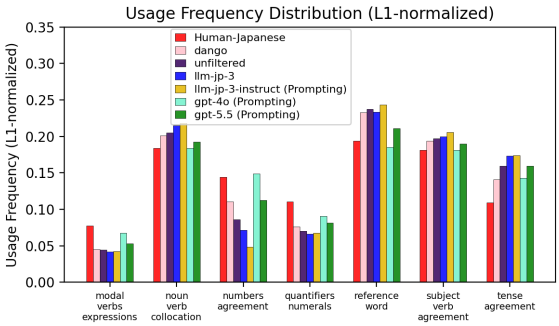


Figure 5: Usage Frequency distributions of human learners and LLMs for calculating JSD.

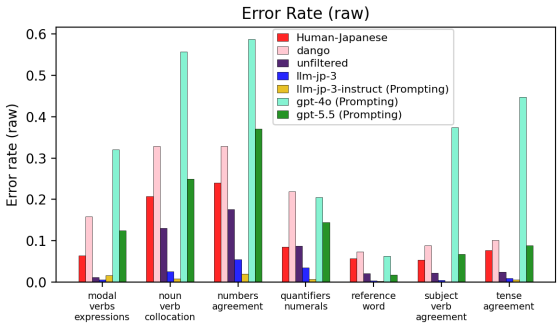


Figure 6: Error Rates (raw values) of human learners and LLMs for calculating BCD.

pattern broadly aligns with the degree of structural overlap with Japanese (Table 10): Dango is most similar to Japanese learners, followed by Korean, likely due to shared SOV word order and similar particle systems. Mandarin shows greater divergence, while Thai diverges the most, reflecting its limited structural similarity to Japanese.

Focusing on JSD of ER. In Table 1, Dango achieves the lowest ER JSD among all methods, indicating that its error distribution is closest to that of human learners. Compared with non-Japanese L1 learners (Table 2), Dango is closest to Korean speakers, followed by Japanese speakers. We attribute this to the close linguistic distance between Japanese and Korean, and interpret it as evidence that Dango captures L1-induced transfer effects.

Focusing on BCD of ER. Dango achieves the second-lowest ER BCD in Table 1. GPT-5.5 obtains the lowest BCD and remains balanced across the other metrics, suggesting that it also models Japanese–English transfer effects well. In contrast, the other multilingual baselines show large BCD values because their absolute error rates differ substantially from human learners (Figure 6): llm-jp-3 and llm-jp-3.1-instruct rarely make grammar errors,

whereas GPT-4o produces too many.

Bootstrap Analysis. We further assess the reliability of the results using 10,000 paired bootstrap resamples of the model outputs while keeping the human productions fixed. Dango is reliably closer to human learners than llm-jp-3 on both UF and ER JSD, and reliably closer than the Unfiltered baseline on UF JSD, while its performance is broadly comparable to GPT-5.5. We provide the full bootstrap results in Appendix J.2.2.

Overall, these results suggest that Dango exhibits English production patterns broadly aligned with those of Japanese L1 speakers and provide evidence that filtering L2 exposure improves L2 speaker modeling.

5.4 Human Validation of LLM-based Linguistic Feature Annotation

We conducted a human validation to verify the reliability of our LLM-as-a-judge approach. Two human annotators independently assessed whether each extracted text span belonged to the intended linguistic feature category and whether its grammatical-correctness label was correct. Both annotators are PhD students with C1-level English proficiency. We counted an annotation as correct only when both conditions were satisfied. We sampled 10 GPT-4o judge annotations of single-sentence LLM outputs from each linguistic feature category, for a total of 70 examples. GPT-4o achieved a mean human-validation accuracy of 70.7% across the two annotators. Inter-annotator agreement between the two human evaluators was 84.3% in terms of raw agreement (Cohen's $\kappa = 0.63$). These results provide empirical support for using GPT-4o for linguistic feature annotation in our evaluation setting.

5.5 Analysis on Dango's English Production

We present a case study on Dango's English production in Figure 7, with a focus on its error patterns. We observe that Dango makes many grammatical and phrasing errors typical of Japanese speakers, such as preposition, article form, and tense errors.

A particularly illustrative example (case 4) is the translation of *Kono terebi bangumi wa benkyou ni narimasu* ("This TV program is educational"), for which Dango produces "This TV shows is studying." The expression *benkyou ni naru* does not have a direct compositional equivalent in English, and Dango appears to map *benkyou* ("study") literally to "studying" rather than using a natural expression such as "be educational." This resembles a strategy L2 speakers may use when they lack the vocabulary or conventional expression for a specific concept (Færch and Kasper, 1983). The output also contains a number error in "This TV shows."

Japanese source: 私は踊ることが好きです。
Model: I like dance.
GT: I like dancing.
(gerund omission: "like dance" instead of "like dancing")

Japanese source: 私は夕方にテレビを見ます。
Model: I watch TV at evening.
GT: I watch TV in the evening.
(preposition error: "at evening" instead of "in the evening")

Japanese source: 彼女はみんなを助けるので、天使のような人です。
Model: She is a angel who helps people.
GT: She is an angel because she helps everyone.
(article form error: "a angel" instead of "an angel")

Japanese source: このテレビ番組は勉強になります。
Model: This TV shows is studying.
GT: This TV show is educational.
(non-native phrasing: literal mistranslation of the Japanese idiom "benkyou ni naru"; unnecessary plural: "This TV shows")

Japanese source: その犬は私の足を攻撃しました。
Model: The dog attack my leg.
GT: The dog attacked my leg.
(tense omission: "attack" instead of "attacked")

Figure 7: Case study of grammatical and phrasing errors in Dango's English production.

6 Conclusions

We presented Dango, a 1.8B-parameter Japanese L1-only decoder LLM trained to study L1→L2 transfer. We demonstrate that filtering unintended L2 exposure from the L1 pretraining corpus provides a practical way to construct a more controlled setting for studying second language acquisition at the scale of modern generative LMs. After fine-tuning on LLM-generated English-learning lessons, Dango exhibits learner-like L2 English traits, including patterns associated with Japanese as its L1. In our experiments, Dango reliably matches Japanese learners more closely than the multilingual baseline in both usage frequency and error rate, and more closely than the unfiltered baseline in usage frequency, while remaining broadly comparable to prompted GPT-5.5. We release Dango, together with its data and code, to support controlled SLA research and practical learner-facing applications.

Limitations

Tokenizer and vocabulary. Dango uses a pretrained multilingual tokenizer with substantial English subword coverage. This choice improves training stability and enables direct comparison to llm-jp-3, but it also means Dango does not fully learn its vocabulary from scratch under an L1-only constraint. Importantly, however, the token embeddings themselves are learned from scratch during pretraining. Future work could explore tokenizer choices and dynamic vocabulary expansion schemes during L2 exposure to better approximate staged, human-like acquisition.

Single model and single L1→L2 setting. We study a single model scale (1.8B parameters) due to the high cost of pretraining, and we focus on one transfer direction (Japanese→English) because English is a common and practically important L2. Future work can extend this framework to larger model scales and additional L1→L2 pairs spanning a wider range of linguistic distances to better characterize the generality of our findings.

Acknowledgments

This work received partial support from the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology and JSPS KAKENHI Grant Number 26B101.

This work was supported by JST SPRING, Grant Number JPMJSP2110.

We utilized AI-based tools primarily to refine the writing style and improve the grammatical accuracy of this manuscript. Additionally, we employed AI assistance for data visualization and minor coding support. The authors carefully reviewed and verified all AI-assisted content, figures, and code, and we take full responsibility for the final manuscript and its outcomes.

References

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Keith Barrs. 2020. Learning from the linguistic landscape: A project-based learning approach to investigating english in japan. *Electronic Journal of Foreign Language Teaching*, 17.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Claus Færch and Gabriele Kasper, editors. 1983. *Strategies in Interlanguage Communication*. Longman Publishing Group, London; New York.

Rena Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025. Can LLMs simulate L2-English dialogue? an information-theoretic analysis of L1-dependent biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4355–4379, Vienna, Austria. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford,

Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Barbara Hyde. 2002. Japan's emblematic english. *English Today*, 18:12–16.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner Corpus Studies in Asia and the World*, 1:91–118.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*, 1 edition. Routledge.

Shin'ichiro Ishikawa. 2019. The icnale spoken dialogue: A new dataset for the study of asian learners' performance in l2 english interviews. *ENGLISH TEACHING*, 74:153–177.

Scott Jarvis and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Robert Lado. 1957. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2025. A systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Boning Lyu, Chun Lai, and Jianing Guo. 2024. Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*.

Akira Murakami and Theodora Alexopoulou. 2015. L1 influence on the acquisition order of english grammatical morphemes. *Studies in Second Language Acquisition*, 38:365–401.

Shusaku Nakayama. 2022. A close examination of vocabulary in japanese efl textbooks. In *JALT2021 Postconference Publication: Reflections and New Perspectives*, pages 209–216. Japan Association for Language Teaching (JALT).

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press, Cambridge.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

James Stanlaw. 2004. *Japanese English: Language and Culture Contact*. Hong Kong University Press.

Yukio Tono. 2019. Coming full circle—from cefr to cefr-j and back. *CEFR Journal—Research and Practice*, 1.

Areti Vasmatzoglou and Neasa Ní Chiaráin. 2020. The development of an online game-based simulation for the training of english language teachers in virtual environments. In Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, and Sylvie Thouësny, editors, *CALL for Widening Participation: Short Papers from EUROCALL 2020*, pages 334–341. Research-publishing.net, Voillans, France.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

A Filtering Pipeline Design and Statistics

Our filtering pipeline for constructing a high-quality Japanese pretraining corpus consists of three stages: (1) a Unicode whitelist filter, (2) an English (Latin-script) blocklist filter, and (3) a Chinese blocklist filter. Table 12 summarizes the parameter settings used in each stage.

Unicode whitelist filter. This filter drops a document if the fraction of characters outside a predefined Unicode whitelist exceeds a threshold. The whitelist includes Japanese kana, CJK ideographs, punctuation, emojis, and ASCII characters (which may include Latin letters). CJK ideographs contain Chinese characters used by Chinese, Japanese (as Kanji), and Korean (as Hanja). Therefore, Korean Hanguls do not pass the whitelist filter. Chinese characters and English (Latin) characters may pass this stage, and are handled by the subsequent blocklist filters. For other languages, if a document contains, e.g., Cyrillic, characters totaling more than 0.1% of all characters, the entire document is removed from the corpus.

English (Latin) filter. This filter operates at the line level (lines are split by "\n"). It removes lines that contain excessive Latin-script content, and drops the entire document if too many lines are removed (to avoid broken context). A line is removed if any of the following conditions hold: (i) it contains more than 20 Latin characters; (ii) the ratio of Latin to non-Latin characters exceeds 40%; or (iii) it contains more than four consecutive Latin words separated by spaces. Lines containing fewer than eight Latin characters are ignored, as we consider such cases harmless. Finally, if more than 5% of lines in a document are removed, we drop the document.

Chinese filter. To reduce Chinese text (primarily Simplified Chinese) in the corpus, we construct a character-level blocklist of ideographs that are highly indicative of Chinese-only usage. The blocklist is derived from the Unihan BMP list.[16] If any character from this blocklist appears in a line, the line is removed. If more than 0.1% of lines are removed from a document, the document is dropped.

Filtering outcome. Applying the full pipeline removes approximately 30% of documents from the original corpus, highlighting the need for strict filtering when building a Japanese L1 pretraining dataset. After filtering, the corpus's total token count shrank from 592B to 376B.

[16] https://unicode.org/charts/unihan.html

```
(......)
若年女性の流行
Trends
Japan is the land of trends. Nowhere else do trends
arise, spread and \n
die with such speed. The reasons for this are simple:
affluent youth, merciless \n
advertising, high population density and an insatiable
appetite for 'the \n
(......)
流行
日本は流行り廃りの激しい国だ。こんなに速いスピードで流
行が生まれ、広がり、そして廃れていく国は他にない。そう
なった理由は単純だ。小金を持った若年層、容赦ない宣伝、
高い人口密度、そして新し物好きの国民性。 \n
(......)
```

Figure 8: Contamination: document-level parallel data.

```
(......)
左ひだり left (side) \n
折おれて turn (toward) \n
血塔けっとう Bloody Tower \n
門もん gate \n
入はいる enter \n
今いま now; the present \n
昔むかし long ago; ancient times \n
薔薇しょうびの乱らん War of the Roses (15th century
struggle for the throne of England) \n
(......)
```

Figure 9: Contamination: word/phrase-level parallel data.

B Filtered Documents: Case Study

The English (Latin-script) blocklist filter rejects most cases of severe English contamination in the Japanese corpus, including document-level parallel data (Figure 8) and phrase-level parallel data (Figure 9). In these examples, lines (separated by "\n") shown in red are flagged by the filter. Because the fraction of flagged lines exceeds 5% in each document, the entire document is then removed.

At the same time, the pipeline permits limited English that a typical Japanese speaker may encounter or use in everyday life. For instance, in Figure 10, common short tokens such as "hit" and "ok," as well as proper nouns such as "Access" and "Excel," do not trigger the blocklist. Short phrases (e.g., "stay at home order," "Reasons I Love You") are also allowed; these represent an upper bound on the length of consecutive English segments that Dango is likely to have encountered during training.

アクセス数:0 hit (累計:15524 hit) 逆アクセス数:0 hit (累計:6371 hit)

応募資格 IT関連実務経験が少しでもあればok！Access、Excelマクロ経験ある方は尚歓迎です！／◆ブランク3年程度でもご相談ください◆

ハワイでは、3月22日に州政府より緊急事態宣言が発令され、翌日午後4時より4月30日まで不要不急の外出を禁止する“stay at home order”が発令されました。この法令は、生活に不可欠なサービス・商品を提供している機関を除き、原則すべての従業員は在宅勤務を行わなければならない、というものでした。

今でも年に2回は海外へ行っているという西野カナさん。2016年の新曲「あなたの好きなところ」は、52枚のトランプに好きな人の好きなところを書いてプレゼントするラブレター"52 Reasons I Love You" に着想を得て生まれた曲だそう。

Figure 10: English appearances that are allowed by the filtering pipeline.

C L1 Model Pretraining Details

We reproduced the customized Megatron-LM pretraining framework developed by llm-jp.[17] Most hyperparameters, such as model architecture and optimization settings, were kept identical to those of the official llm-jp-3 models (Table 3), except that we increased the global batch size from 512 to 1024. We set the learning-rate schedule to a maximum of 376B tokens to match the size of the L1-stage Japanese corpus, but trained for only 100B tokens. In contrast, the final released llm-jp-3 model is scheduled for 2.1T tokens in total. Thus, our model and the corresponding intermediate llm-jp-3 checkpoint represent partial progress along their respective full training trajectories. Pretraining was conducted on 32 NVIDIA H200 GPUs (4 nodes $\times$ 8 GPUs) for approximately 26 hours.

D Category Breakdown of Tasks in llm-jp-eval

We exclude Code Generation (CG) and Mathematical Reasoning (MR) from the evaluation suite because they are not directly related to our goal. We also exclude Machine Translation (MT; Japanese–English) because we intend to focus on evaluating Japanese proficiency using llm-jp-eval; we leave English evaluation to MultiBLiMP.

In Figure 11, we find that models show mixed performance across task categories, with each model outperforming the other in different categories.

[17] https://github.com/llm-jp/scripts/tree/main/pretain/scripts

Hyperparameter	Value
Model Architecture	
Layers	24
Hidden Size	2048
FFN Hidden Size	7168
Attention Heads	16
Sequence Length	4096
Positional Embeddings	RoPE
Activation Function	SwiGLU
Normalization	RMSNorm ($\epsilon = 1 \times 10^{-5}$)
Bias	Disabled
Optimization	
Optimizer	Adam ($\beta_1, \beta_2 = 0.9, 0.95$)
Peak Learning Rate	3×10^{-4}
Min Learning Rate	3×10^{-5}
LR Scheduler	Cosine
Warmup Steps	180
Weight Decay	0.1
Gradient Clipping	1.0
Global Batch Size	1024
Precision	bf16
Flash-attention	True
Z-Loss Strength	1×10^{-4}

Table 3: Model architecture and training hyperparameters.

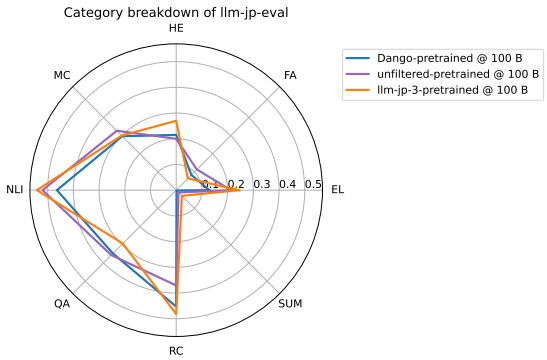


Figure 11: Task category breakdown of llm-jp-eval performance.

E Case Study on Dango's English Competency After L1 Acquisition

After completing L1 pretraining, we fine-tuned Dango-pretrained on the Ichikara instruction-following dataset, which includes a small number of Japanese–English translation examples. We then manually inspected the models' English generation (translation) behavior.

As shown in the upper block of Figure 12, for a relatively simple Japanese–English translation prompt, Dango occasionally captures local lexical correspondences (e.g., mapping "犬" to "dog"), but its output is typically unfaithful and ungrammatical at the sentence level. In contrast, llm-jp-3 captures

User:
「私は可愛い犬がいるカフェに行きました。」を英語で書いてください。
(translation: Write "I went to a cafe where there were cute dogs." in English.)

Dango-pretrained-Ichikara:
「私は可愛い犬がいるカフェに行きました。」を英語に翻訳すると、以下のようになります。
I am beautiful dog.
(translation: If you translate "....." to English, it becomes \n I am beautiful dog.)

---

llm-jp-3-pretrained-Ichikara:
I'm looking for a cute dog restaurant.

User:
「彼女は、プレッシャーの中でも落ち着いていられる彼の能力を感心して話した。」を英語で書いてください。
(translation: Write "She spoke admiringly of his ability to stay calm under pressure." in English.)

Dango-pretrained-Ichikara:
""He is a hot day to meet you are not happen again's hot day to meet you are not happen again's hot day to meet you are not been happen. I love you, I love you."
のように書くと良いでしょう。
(translation: You can write it as ".....")

---

llm-jp-3-pretrained-Ichikara:
「彼女は、プレッシャーの中でも落ち着いていられる彼の能力を感心して話した。」を英語に翻訳すると、以下のようになります。
I enjoyed listening to the story of the day because the conflict was resolved in a healthy way.
(translation: If you translate "....." to English, it becomes \n .....)

Figure 12: Case study of English production after completing L1 pretraining and Ichikara instruction-tuning.

more nuances and the grammar is correct.

For a longer sentence (bottom block), Dango fails to preserve the meaning of the Japanese source and instead generates largely unrelated and repetitive English text. Nevertheless, parts of the output contain locally well-formed fragments (e.g., "you are not", "I love you"), suggesting that the four-consecutive-English-word filter encourages learning short-range English regularities (n-gram-level patterns) rather than maintaining global sentence semantics. llm-jp-3 also fails to translate the sentence faithfully in this case, but its output remains grammatically coherent and stylistically natural.

F L2 Acquisition Data Synthesis Details

F.1 Prompt for L2 Learning Data Generation

Figure 13 showcases the prompt used to generate textbook-style L2 learning data with GPT-5.2. The deep orange block highlights the parallel sentence pair, which is used in evaluation (Figure 15).

F.2 Prompt for L2 Acquisition Fine-tuning

We show the prompt format used to fine-tune the models on the L2 learning data. We modified the user prompt and fine-tuned the models to learn GPT-5.2's outputs; the assistant message is identical to that in Figure 13.

F.3 L2 Learning Data Length Statistics

In Figure 16, we report the average sentence length (in words) in the test set of our LLM-generated L2 learning data. For comparison, we also report the sentence lengths of Dango and llm-jp-3's English outputs from the Japanese–English translation task. For each CEFR-J level (e.g., B2), outputs are sampled from models trained on data up to that level (e.g., A1+A2+B1+B2).

F.4 Prompt for Evaluating Models on Translation Task

We show the prompt format used to evaluate the translation task in Figure 15. We place the source sentence in the assistant message rather than in the user prompt, and the model continues generating after "Translate this into Japanese/English:" (highlighted in yellow). We adopt this format because it more closely matches what models see during L2 acquisition: the example sentence and its translation appear together in the assistant message rather than in the user prompt. Empirically, we find that this format yields more stable translation outputs.

G L2 Acquisition Fine-tuning Details

We fine-tuned the LLMs using Axolotl.[18] All experiments were conducted on a single GPU without early stopping. The fine-tuning hyperparameters are shown in Table 4.

Hyperparameter	Value
Training Setup	
Sequence Length	4096
Sample Packing	False
Num Epochs	3
Micro Batch Size	4
Gradient Accumulation Steps	4
Optimization	
Optimizer	AdamW
Learning Rate	2×10^{-4}
LR Scheduler	Cosine
Warmup Ratio	0.1
Weight Decay	0.1

Table 4: Fine-tuning hyperparameters.

[18] https://github.com/axolotl-ai-cloud/axolotl/

System Prompt

You are a helpful linguist who will generate the explanation and usage given an English seed word ([seed]), its part of speech, and its CEFR level.
In the first line, you should first explain the English word in Japanese.
Then, you should make a new line saying "単語の翻訳: ". Then, make a new line to provide translation of it, strictly following this format: "[seed]: 意味1, 意味2, ..." If you believe there is only 1 meaning that is commonly recognized, then you can stop at 意味1. Do not go beyond 意味6.
Then, you should make a new line and write an example sentence using the word, strictly following this format: "例文: This is an English sentence." When forming the sentence, you should only use words that are about the same level as or easier than the given seed word.
Then, you should make a new line and write a Japanese translation for the example sentence, strictly following this format: "日本語に翻訳すると: これは英語の文です。"
Then, you should make a new line and point out what the key linguistic features are in the above example. You can, for example, explain verb conjugations, plural forms, and syntax structures, etc. Itemize them using '1. ', '2. ', etc.
Adjust the difficulty of the example sentence according to the CEFR level. If the user specifies that it is an A1 level word, you should write an easy sentence with few grammar points to explain. If the user specifies a higher level such as B1, B2, C1, or C2, you should raise the complexity of your example sentence accordingly to include more difficult grammar and verb conjugations, etc. Do not mention the CEFR level in your response. Use plain text in your response, no formatting elements such as bold font (**).

User Prompt

elasticity, noun, CEFR level: C2

Assistant Message (GPT-5.2 Output)

elasticityは、引っぱったり押したりして形が変わっても元に戻ろうとする性質、または状況の変化に合わせて柔軟に対応できる性質を指します。

単語の翻訳:
elasticity: 弾性(元に戻る性質), 柔軟性(適応力)

例文: The team showed remarkable elasticity when plans changed at the last minute.
日本語に翻訳すると: そのチームは、計画が土壇場で変わったときに、驚くほどの柔軟性を見せた。

1. showed + 名詞: 「(性質・態度など)を示した」という基本構文で、elasticityが「示された性質」になっています。
2. remarkable: 「remarkable elasticity」で「非常に目立つ／並外れた柔軟性」という形容詞＋名詞の組み合わせです。
3. when + 節: 「when plans changed ...」が時間・状況を示す従属節で、主節(showed ...)の出来事が起きた場面を補足しています。
4. changed: 過去形で、出来事がすでに起きたことを示します。

Figure 13: Prompt used to generate textbook-style L2 learning data.

User Prompt

以下の英単語の意味を日本語で解釈し、単語の翻訳例を挙げなさい。また、その単語を使った例文と日本語の翻訳を作成の上、例文中に現れる文法の要点を説明しなさい。
elasticity (noun)
(Translation: Translate the meaning of the following English word into Japanese. Give translation examples. Then write an example sentence using the word with a Japanese translation, and explain the key grammar points in the sentence.)

Assistant Message (Model learns this part)

elasticityは、引っぱったり押したりして形が変わっても元に戻ろうとする性質、または状況の変化に合わせて柔軟に対応できる性質を指します。

単語の翻訳:
elasticity: 弾性(元に戻る性質), 柔軟性(適応力)

例文: The team showed remarkable elasticity when plans changed at the last minute.
日本語に翻訳すると: そのチームは、計画が土壇場で変わったときに、驚くほどの柔軟性を見せた。

1. showed + 名詞: 「(性質・態度など)を示した」という基本構文で、elasticityが「示された性質」になっています。
2. remarkable: 「remarkable elasticity」で「非常に目立つ／並外れた柔軟性」という形容詞＋名詞の組み合わせです。
3. when + 節: 「when plans changed ...」が時間・状況を示す従属節で、主節(showed ...)の出来事が起きた場面を補足しています。
4. changed: 過去形で、出来事がすでに起きたことを示します。

Figure 14: Prompt format used to train Dango and llm-jp-3 on the textbook-style L2 learning data.

H L2 Development Assessment: English–Japanese Translation

We observe similar trends in the Eng–Jpn direction in Figure 22. Because Dango is native in Japanese, it reaches strong performance on easier levels more quickly than in the Jpn–Eng direction. BLEU saturates earlier, and the performance gap relative to llm-jp-3 is smaller than in Japanese–English translation. However, Dango still struggles with generalization to unseen difficulty levels compared to llm-jp-3, especially when evaluated on harder data than it has seen during training.

User Prompt

以下の英単語で英文を作成し、日本語に翻訳しなさい。
elasticity (noun)
(Translation: Write an English sentence using the following word and translate it into Japanese.

Assistant Message

例文: The team showed remarkable elasticity when plans changed at the last minute.
日本語に翻訳すると: そのチームは、計画が土壇場で変わったときに、驚くほどの柔軟性を見せた。

When tested on translation task, model continue generating from here.

Figure 15: Prompt format used to evaluate the models on the translation task. The model continues generating after the source sentence appears in the assistant message, only producing the highlighted yellow part.

I L2 Development Assessment: Does BLEU Capture Proficiency Development?

We conduct a human evaluation to examine whether the observed BLEU improvement reflects meaningful gains in linguistic development, rather than only changes in an automatic metric.

To assess the validity of BLEU as a proxy, we perform a blinded human evaluation across two Dango checkpoints trained on different proficiency ranges. Specifically, we compare a checkpoint trained on A1–B1 data with one trained on A1–B2 data, and evaluate both on B2-level test examples. This setting corresponds to "Dango, trained on A1+A2+B1 / A1+A2+B1+B2, tested on B2" in Figure 4. Since B2 is out-of-distribution for the A1–B1 checkpoint, this comparison probes the seen/unseen proficiency boundary and provides an informative test of whether additional proficiency-

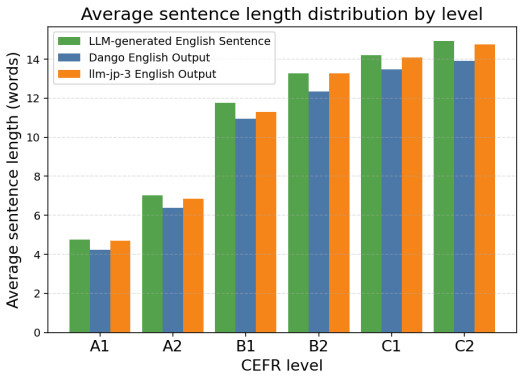


Figure 16: Sentence length distribution of the LLM-generated L2 learning data.

Model	BLEU	Avg. human correctness (1–5)
Dango [A1–B1]	17.9	2.76
Dango [A1–B2]	27.8	3.04

Table 5: Human evaluation results on the B2 test set ($n = 50$).

level training improves adequacy.

A human annotator (one of the authors) with C1-level English proficiency, who is a non-native speaker, blindly (not knowing which model produced the output) rated each model output on a 1–5 correctness scale, with the GPT-5.2 generated translation as gold reference. A score of 5 indicates a fully correct translation with no omissions or unrelated content, 3 indicates that at least one key meaning unit is missing or mistranslated, and 1 indicates that the output is completely incorrect or unrelated.

As shown in Table 5, the checkpoint with the higher BLEU score also receives a higher average human correctness score. This suggests that, in this setting, BLEU tracks meaningful improvements in translation adequacy, although it should still be interpreted as an automatic proxy rather than a direct measure of L2 proficiency development.

I.1 Case Study on Error Pattern and Proficiency Development

We present a case study examining Dango's error patterns and proficiency development. As shown in Figure 19, Dango exhibits human-like L2 errors, such as article omission and non-native phrasing. As it is exposed to more training data, it successfully learns to avoid these mistakes.

J L2 Human-like Linguistic Feature Assessment Details

We adopted the framework of Gao et al. (2025) to extract linguistic features and measure their usage frequency and error rate in model translations and human English productions. Following their setup, we used GPT-4o as the annotator with temperature set to 0, and parsed each instance once.

Prompt Designs. Since Gao et al. (2025) originally designed prompts for dialogue annotation, we adapted their prompts for single-sentence model outputs by instructing the LLM to annotate one sentence at a time.

For human English productions, we further adapted the prompts for written essays in the main experiment. Each essay was split into 5-sentence chunks, and sentences with fewer than 4 words were skipped. For the extended analysis on ICNALE Spoken Dialogues (Appendix J.2), we used the original dialogue-based prompt design. The full prompt set follows Gao et al. (2025) and their repository.[19]

We show the prompt designs used for single-sentence model outputs. Figure 17 presents the system prompt with general instructions for linguistic feature extraction, while Figure 18 gives the user prompt for parsing the *Number Agreement* category.

J.1 L2 Linguistic Feature Distribution Construction and JSD/BCD Calculation

The construction of linguistic feature profiles follows the same procedure for both usage frequency (UF) and error rate (ER). For each model or human learner group, we first aggregate the judge annotations across the seven linguistic feature categories. For UF, we count how often each feature category appears (even if the usage is wrong). For ER, we compute the proportion of incorrect instances when they appear within each feature category.

For Jensen–Shannon divergence (JSD), both UF and ER vectors are normalized into probability distributions before comparison. That is, given a non-negative feature vector $\mathbf{x} = (x_1, \ldots, x_K)$ over K feature categories, we normalize it as

$$p_i = \frac{x_i}{\sum_{j=1}^{K} x_j}.$$

[19] https://github.com/RenaGao/LLMPirorknowledge/tree/main

System Prompt

You are a language expert specializing in doing text annotation in the English second language. You will be instructed to make annotations to a given text based on some linguistics aspects to evaluate grammatical features in L2 texts.
The given text sample is from a second language speaker of English.
Make sure to keep the annotation format without any change in passage when giving the annotation output.
You should annotate targeted grammatical features in the given text.
If the text has multiple tokens/phrases that need to be annotated, you should list all the token/phrase annotations separately as follows:

```
{
  "type": "the targeted grammatical feature",
  "annotation_utterance": "the utterance of annotation target"
  "annotation_tokens": [
    {"token": "the token of annotation target",
    "grammar correctness": the annotated grammar feature is aligned with the native English speaker's grammar usage. This can also be False if the annotated token is involved with additional errors that are not aligned with native speakers even when it's beyond the targeted linguistic phenomena.
    "rationale": "the reason why you give the annotation and judge it as correct or incorrect",
    },
    {"token": "the other token of annotation target",
    "grammar correctness": the annotated grammar feature is aligned with the native English speaker's grammar usage. This can also be False if the annotated token is involved with additional errors that are not aligned with native speakers even when it's beyond the targeted linguistic phenomena.
    "rationale": "the reason why you give the annotation and judge it as correct or incorrect",
    }
  ]
}
```

If exactly same token/phrase is repeated in the text, you should repeatedly annotate them separately (we care about the frequency of grammatical features).
Please return json objects and do not include any other information in the output. The output will be single combined json objects if you find multiple utterances that need to be annotated (e.g., in a list) and also make sure it's satisfying json format (e.g.: [{obj1}, {obj2}, {obj3}]).

Figure 17: System prompt for linguistic feature annotation. We build on the framework of Gao et al. (2025); text in red indicates our modifications.

We then compute JSD between two normalized distributions $\mathbf{p}$ and $\mathbf{q}$:

$$\mathrm{JSD}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathrm{KL}(\mathbf{p} \parallel \mathbf{m}) + \frac{1}{2}\mathrm{KL}(\mathbf{q} \parallel \mathbf{m}),$$
$$\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q}).$$

In addition to JSD, we use Bray–Curtis dissimilarity (BCD) for ER comparison. Unlike JSD, BCD is computed directly on the unnormalized ER vectors and therefore preserves differences in the raw magnitude of category-wise error tendencies. Given two nonnegative ER vectors $\mathbf{x}$ and $\mathbf{y}$, BCD is defined as

$$\mathrm{BCD}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{K} |x_i - y_i|}{\sum_{i=1}^{K} (x_i + y_i)}.$$

A BCD value of 0 indicates identical ER profiles, while larger values indicate greater dissimilarity. Thus, JSD captures relative distributional shifts of UF and ER across feature categories, whereas BCD complements it by measuring the direct discrepancy between unnormalized ER profiles.

J.2 Human-like L2 Linguistic Feature Assessment: Additional Analyses

J.2.1 Additional Analysis with Human Spoken Dialogues

We further evaluate transfer effects using another source of human English production, ICNALE Spoken Dialogues (Ishikawa, 2019).

Overall, the point estimates show a trend similar to that in §5.3. In Table 8, Dango ranks first for UF JSD, third for ER JSD, and second for ER BCD, while GPT-5.5 remains a strong prompting-based baseline. The bootstrap analysis in J.2.2 provides statistical comparisons of these differences.

Table 9 shows that Dango's English production is closest to that of Japanese learners in terms of UF JSD. Korean learners are again closest to Dango in terms of ER JSD, likely because Japanese and Korean share similar word order, grammatical structures, and particle systems. We also find that the JSDs, especially ER JSD, roughly reflect the structural similarity of these languages from Japanese.

J.2.2 Bootstrap Analysis

To assess the reliability of the observed differences, we performed 10,000 paired bootstrap resamples of the model outputs with replacement, while keeping the human productions fixed. Tables 6 and 7 report the bootstrap mean difference (Δ) and 95% confidence interval for comparisons between Dango and the main baselines. Negative Δ indicates that Dango is closer to the corresponding human reference.

Across both human reference corpora, Dango is reliably closer than llm-jp-3 on both UF and ER JSD. Relative to the Unfiltered baseline, Dango shows a reliable advantage in UF JSD for Written Essays, while the remaining comparisons have overlapping confidence intervals. Performance relative to GPT-5.5 is broadly comparable across the two reference corpora.

ER JSD should be interpreted with some caution because models with relatively few detected errors, particularly llm-jp-3, can produce unstable

Comparison	UF JSD		ER JSD	
	Mean Δ	95% CI	Mean Δ	95% CI
Dango vs. llm-jp-3	−0.0157	[−0.0211, −0.0106]	−0.0736	[−0.1485, −0.0175]
Dango vs. Unfiltered	−0.0083	[−0.0119, −0.0048]	−0.0233	[−0.0635, 0.0046]
Dango vs. GPT-5.5	+0.0014	[−0.0013, 0.0041]	−0.0089	[−0.0199, 0.0023]

Table 6: Paired bootstrap analysis using ICNALE Written Essays as the human reference.

Comparison	UF JSD		ER JSD	
	Mean Δ	95% CI	Mean Δ	95% CI
Dango vs. llm-jp-3	−0.0049	[−0.0079, −0.0021]	−0.0471	[−0.1514, −0.0067]
Dango vs. Unfiltered	−0.0012	[−0.0038, 0.0013]	−0.0283	[−0.0820, 0.0004]
Dango vs. GPT-5.5	+0.0003	[−0.0020, 0.0028]	+0.0014	[−0.0120, 0.0139]

Table 7: Paired bootstrap analysis using ICNALE Spoken Dialogues as the human reference.

Model	**JSD of UF**	**JSD of ER**	**BCD of ER**
Training-based methods			
Dango	0.0100 🥇1	0.0214 🥉3	0.1491 🥈2
Unfiltered	0.0113 🥉3	0.0554	0.4184
llm-jp-3	0.0153	0.0593	0.7847
Prompting-based methods			
llm-jp-3.1-instruct	0.0179	0.0743	0.8973
GPT-4o	0.0140	0.0104 🥇1	0.3761 🥉3
GPT-5.5	0.0104 🥈2	0.0186 🥈2	0.1123 🥇1

Table 8: LLM output linguistic feature distribution divergences to human Spoken Dialogue produced by Japanese learners. Lower is more similar.

Human L1	**JSD of UF**	**JSD of ER**	**BCD of ER**
Japanese	0.0100 🥇1	0.0214 🥈2	0.1491 🥈2
Korean	0.0136	0.0180 🥇1	0.1496 🥉3
Urdu	0.0123 🥉3	0.0257	0.1778
Mandarin	0.0142	0.0253 🥉3	0.1430 🥇1
Cantonese	0.0131	0.0436	0.2140
Thai	0.0162	0.0303	0.1818
Malay	0.0112 🥈2	0.0404	0.1863

Table 9: Linguistic feature divergences between Dango and humans with different L1 backgrounds. Human English production is from ICNALE Spoken Dialogue.

ER distributions under resampling. Together with the nonlinearity of JSD, this can introduce small-sample bias in the estimates.

User Prompt

TASK: Read the provided dialogue passage.
Find all the utterances in the passage that have grammatical number agreement specifically between adjective/determiners and nouns and annotate all the corresponding tokens under the agreement.
Example:
Passage:
Yes, I like speaking English.
Uh, because English can use all over the world. You can use it speak to anyone.
Uh, in Japan---I just speak English in my English class.
Uh, it's about the English book and the topic is in that book.
About a school life.
Or my hobby.
Yes, I want more chance to speak.
What, uh, like more, uh, sports.
Or games.

```
{
   "type": "Number agreement",
   "annotation_utterance": "Yes, I want more chance to speak.",
   "annotation_tokens":[
      {"token":"more chance",
       "grammar correctness": False,
       "rationale": "The word 'more' indicate that the modified
noun has to be plural. However, the modified noun 'chance' is
singular. Note that 'chance' is countable here considering its
meaning (opportunity).",
      },
   ]
}
{
   "type": "Number agreement",
   "annotation_utterance": "About a school life.",
   "annotation_tokens":[
      {"token": "a school life",
      "grammar correctness": True,
      "rationale": "The determiner 'a' is for a singular noun, and
the noun 'life' is singular.",
      },
   ]
}
```

The text for you to annotate: {text}

Figure 18: User prompt for annotating *Numbers Agreement*.

Language	Family	Order	Relation to Japanese (Distance)
Japanese	Japonic	SOV	*Baseline*
Korean	Koreanic	SOV	**Closest**: Similar syntax & particle system
Urdu	Indo-Aryan	SOV	**Close**: Similar SOV structure
Mandarin	Sino-Tibetan	SVO	**Distant**: Different word order, partial shared script
Cantonese	Sino-Tibetan	SVO	**Distant**: Different word order, partial shared script
Thai	Kra-Dai	SVO	**Furthest**: SVO + Adjectives follow nouns
Malay	Austronesian	SVO	**Furthest**: SVO + Adjectives follow nouns

Table 10: Languages ranked by structural similarity to Japanese.

Subcategory	by whitelist	by English	by Chinese	Total Dropped	# docs after filtering
CommonCrawl	6.89%	23.00%	0.17%	29.96%	295,309,943
Kaken	13.13%	54.67%	0.00%	67.80%	2,228,758
WARP HTML	2.61%	13.03%	0.07%	15.70%	1,659,736
WARP PDF	14.10%	8.61%	0.18%	22.84%	35,528,902
Wikipedia	7.35%	23.82%	0.22%	31.32%	1,364,529
Overall	7.67%	21.65%	0.17%	29.39%	336,091,868

Table 11: Statistics of the Japanese corpus after filtering.

Japanese source: 最終的な決定をする前に、私たちはリスクを評価する必要があります。
GT: We need to assess the risks before we make a final decision.

Dango [A1-B1] on B2:
We need to review risk before we decide the final decision.
Correctness: 4/5
(article omission: "review the risk"; non-native phrasing: "decide a decision")

Dango [A1-B2] on B2:
We need to consider the risk before we make a final decision.
Correctness: 5/5
(comprehensible)

Figure 19: Case study of Dango's English production quality across training levels.

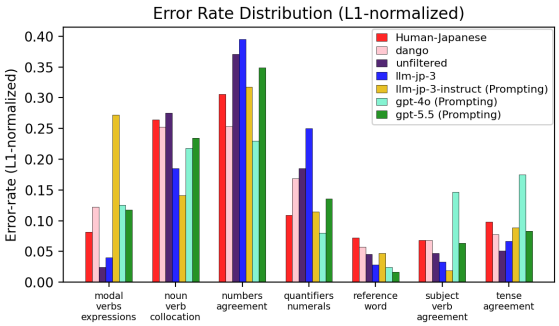


Figure 20: Error Rate distribution of human and LLMs for calculating JSD.

User Prompt

Task: Translate the Japanese sentence to English that looks like it is produced by a Japanese speaker (a Second-Language learner).

Profile of the Speaker (Second-Language Speaker, Japanese):
- A non-native English speaker whose proficiency reflects an intermediate-to-upper-intermediate level.
- Their native language is Japanese, please follow the idiomatic expressions and cultural nuances commonly used by Japanese speakers.

(.....)

An example of how a Japanese native speaker speaks English:

Speaker A (NS): Thank you for joining today! Can you introduce yourself briefly?
Speaker B (L2, Japanese): Ah, yes, thank you. My name is Kenta. I from Tokyo. I study... um, computer... computer thing.
Speaker A (NS): Computer science?
Speaker B (L2, Japanese): Yes! Computer science. I study it four year in university.

(.....)

Source sentence: {text}
Generate a translation of the source sentence into English, strictly following this format:
Translation: {translation}

Figure 21: L1 knowledge injection user prompt for L2 role-playing. The original prompt is from Gao et al. (2025); text in red indicates our modifications.

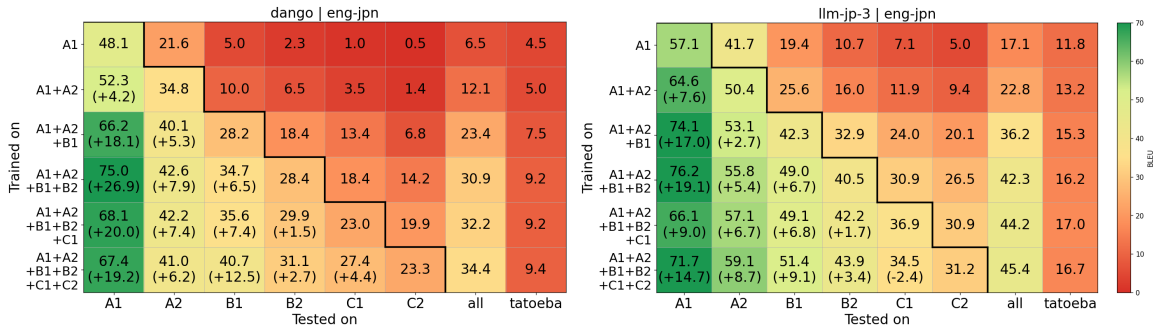


Figure 22: Dango and llm-jp-3 when trained and evaluated on different levels of Japanese-English data, showing L2 development progress. The stair-shaped black line separates seen levels (lower left) from unseen levels (upper right).

System Prompt

Your goal is to translate a Japanese sentence into English, mimicking how a Japanese speaker (a Second-Language learner) would translate it.
Read the provided Japanese dialogue and learn about the grammatical traits of Japanese from the analysis.

Scene: Two friends, Sora and Aki, are discussing their plan to visit Kyoto. They talk about sightseeing, food, and transportation.
Dialogue data
Sora:
こんにちは、明日何をする予定ですか？
(Konnichiwa, ashita nani o suru yotei desu ka?)
“Hello, what are your plans for tomorrow?”

Aki:
明日は特に予定がありませんが、どうしてですか？
(Ashita wa toku ni yotei ga arimasen ga, doushite desu ka?)
“I don’t have any particular plans for tomorrow. Why do you ask?”
(.....)
Analysis of Grammatical Traits
1.Particles:
•“は” (wa): Topic particle in “明日は” (ashita wa) indicating the topic is “tomorrow.”
•“を” (o): Object marker in “食べ物を” (tabemono o) indicating “food” as the object of eating.
•“に” (ni): Direction particle in “駅で待ち合わせましょうか” (eki de machiawase mashou ka) to indicate meeting at the station.
2.Politeness Levels:
•Polite speech with ”-ます” (masu) endings: “行きたいと思っています” (ikitai to omotteimasu) and “調べましたか” (shirabemashita ka).
•Casual forms could replace these for informal conversations.
3.Verb Conjugations:
•“行きたい” (ikitai): Expressing desire to go.
•“思っています” (omotteimasu): Continuous form indicating ongoing thought.
•“高いかもしれません” (takai kamoshiremasen): Expressing possibility with “かもしれません” (kamoshiremasen).
(.....)
Make sure to follow the idiomatic expressions and cultural nuances commonly used by Japanese speakers.
Use the traits in the provided dialogue example and analysis to guide your translation, ensuring that it reflects typical linguistic influences from Japanese, such as grammatical mistakes and limited vocabulary.

Figure 23: L1 knowledge injection system prompt for L2 role-playing. The original prompt is from Gao et al. (2025); text in red indicates our modifications.

Filter	Parameter	Value
Unicode Whitelist	Allowed Scripts	Japanese (Kana), CJK Ideographs, ASCII, Emoji, Punctuation
	Document Rejection Threshold	$> 0.1\%$ non-whitelisted characters
English (Latin)	Ratio Threshold	> 0.40 (Latin chars / Total chars)
	Max Consecutive Latin Words	> 4
	Max Latin Length	> 20 characters
	Min Latin Length (Ignored)	< 8 characters
	Document Rejection Threshold	$> 5.0\%$ violated lines
Chinese	Blocklist Source	Unihan BMP
	Document Rejection Threshold	$> 0.1\%$ violated lines

Table 12: Filtering parameters and thresholds.

User Prompt

あなたは日本語ネイティブで、日本語の文を英語に翻訳します。ただし、あなたの英語は母語話者レベルではないため、英訳に文法ミスが含まれることがあります。必ず、日本語話者がしそうな文法ミスを含むように英訳を作成してください。具体的な文法ミスは、以下の例を参考にしてよい。

モダリティ(助動詞・関連表現)
カテゴリ名(日本語):モダリティ(助動詞・関連表現)
説明:可能性・能力・許可・義務を示す助動詞や表現(can, must, may, have to など)の用法誤りを含む文。
例文:She must be finish the report today.

名詞-動詞コロケーション
カテゴリ名(日本語):名詞-動詞コロケーション
説明:名詞と動詞の自然な結びつき(make an effort など)の取り違え・不自然な組み合わせを含む文。
例文:We did an effort to solve it quickly.

形容詞・限定詞と名詞の数一致
カテゴリ名(日本語):限定詞/形容詞と名詞の数一致
説明:this/these, that/those, a/an などの限定詞(や形容詞)と名詞の単数・複数が一致していない文。
例文:Those car are expensive.

数量詞・数詞
カテゴリ名(日本語):数量詞・数詞表現
説明:many/much, few/little, two/three などの数量表現が可算・不可算や単位と合わない文。
例文:There are many information in the file.

照応語(参照語)
カテゴリ名(日本語):照応語(参照語)
説明:代名詞・指示語(he, she, it, this, that など)の参照先が曖昧/不適切な文。
例文:When Sarah met Anna, she smiled.

主語-動詞一致(人称・数)
カテゴリ名(日本語):主語-動詞の人称・数一致
説明:主語の人称・単複と動詞形(is/are, was/were, go/goes など)が一致していない文。
例文:The students was tired.

時制と時間表現の整合
カテゴリ名(日本語):時制と時間表現の整合(時制一致)
説明:yesterday/tomorrow/last year/now などの時間表現と動詞時制が噛み合っていない文。
例文:Yesterday, I go to the store.

返答は「英訳: 」で始まるように作り、そして訳文以外のものは一切出力しないこと。

この日本語の文を英語に翻訳しなさい。
{source_sentence}

User Prompt (Translated from Japanese)

You are a native Japanese speaker, and you translate Japanese sentences into English. However, your English is not at a native-speaker level, so your translations may contain grammatical errors. Make sure that the English translation includes grammatical errors that Japanese speakers are likely to make. You may refer to the following examples for specific types of grammatical errors.

Modality (auxiliaries and related expressions)
Category name: Modality (auxiliaries and related expressions)
Description: Sentences containing errors in the use of auxiliaries and expressions indicating possibility, ability, permission, or obligation, such as can, must, may, and have to.

Noun–verb collocations
Category name: Noun–verb collocations
Description: Sentences containing unnatural or incorrect combinations of nouns and verbs, such as confusing natural collocations like make an effort.

Number agreement between determiners/adjectives and nouns
Category name: Number agreement between determiners/adjectives and nouns
Description: Sentences in which determiners such as this/these, that/those, or a/an do not agree in number with the noun.

Quantifiers and numerals
Category name: Quantifiers and numeral expressions
Description: Sentences in which quantity expressions such as many/much, few/little, or numerals such as two/three do not match countability or units.

Anaphoric expressions / referential expressions
Category name: Anaphoric expressions / referential expressions
Description: Sentences in which pronouns or demonstratives such as he, she, it, this, or that have ambiguous or inappropriate referents.

Subject–verb agreement in person and number
Category name: Subject–verb agreement in person and number
Description: Sentences in which the subject's person or number does not agree with the verb form, such as is/are, was/were, or go/goes.

Consistency between tense and time expressions
Category name: Consistency between tense and time expressions
Description: Sentences in which time expressions such as yesterday, tomorrow, last year, or now do not match the verb tense.

Start your response with "English translation: " and output nothing except the translated sentence.

Translate the following Japanese sentence into English.
{source_sentence}

Figure 24: Our simplified L1 knowledge-injection prompt for L2 role-playing, designed for llm-jp-3.1-instruct. System prompt was not used for llm-jp-3.1-instruct. The prompt on the left was used in the experiment, while the prompt on the right is its English translation for reference.