Borrowed Words, Borrowed Minds: Probing LLM Choice of English-Derived Loanwords in Japanese

Joseph James
Department of Computer Science, The University of Sheffield
Sheffield, United Kingdom
jhfjames1@sheffield.ac.uk

Abstract

The choice between English-derived loanwords (*gairaigo*) and native Japanese equivalents is a socially meaningful aspect of language use, carrying implications for register, style, and pragmatic interpretation. We introduce a controlled evaluation dataset probing how large language models encode this form of sociolinguistic variation. The dataset comprises 113 interchangeable lexical pairs embedded across six communicative contexts spanning formal and informal, spoken and written registers. We evaluate 16 Japanese-capable LLMs across three complementary tasks: sentence rating, pairwise choice, and masked word prediction. Although both lexical forms were generally rated as natural, models diverged substantially in their contextual sensitivity and lexical preferences, revealing architectural differences in how socially grounded lexical alternatives are represented. These findings suggest that surface fluency may mask instability in modeling pragmatic variation, with implications for socially aware language generation and evaluation. Dataset and prompts will be made publicly available upon publication to facilitate replication and further research.

1 Introduction

The Japanese lexicon is characterised by extensive borrowing, particularly from English (Irwin, 2011; Tomoda, 1999). English-derived loanwords, known as *gairaigo*, coexist with native (*wago*) and Sino-Japanese (*kango*) equivalents and are typically written in katakana. In many cases, loanwords overlap semantically with existing Japanese terms. For example, *henji* (返事) and *ripurai* (リプライ) both denote a reply, yet differ in communicative association and contextual framing. *Henji* functions as a broad native term used across a wide range of settings, whereas *ripurai* is more closely associated with digital communication and contemporary discourse. Such alternations are not neutral substitutions but carry contextual and social meaning.

Loanword usage in Japanese reflects multiple sociolinguistic dimensions. English-derived forms may signal modernity, international orientation, technological currency, or commercial branding, while native equivalents may evoke institutional authority, convention, or cultural continuity. At the same time, register distinctions in Japanese remain highly codified and play a central role in shaping lexical, grammatical, and pragmatic choices in both speech and writing (Liu and Allen, 2014; Dunn, 1999; Matsumoto, 1988). Formal and informal contexts often impose systematic constraints on lexical selection, making register a salient and empirically tractable dimension of sociolinguistic variation. Although lexical alternation cannot be reduced to formality alone, the formal–informal contrast provides a principled baseline for examining contextual sensitivity.

These issues are increasingly relevant in the context of large language models. Contemporary LLMs trained on large-scale Japanese corpora are exposed to diverse registers and communicative styles (Kuribayashi et al., 2021). However, exposure does not necessarily imply appropriate contextual differentiation. Models may generate fluent output while failing to distinguish subtle register constraints or socially appropriate lexical choices. For applications in education, translation, and writing assistance, such distinctions are consequential. If models treat near-equivalent loanword and native forms as interchangeable across contexts, they risk obscuring sociolinguistic nuance.

Despite growing interest in stylistic control and sociolinguistic evaluation of LLMs, systematic examination of context-sensitive lexical alternation in Japanese remains limited. In this paper, we investigate how LLMs select between English-derived loanwords and native Japanese equivalents across structured communicative settings. We introduce a

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 22–36

dataset of 113 interchangeable lexical pairs embedded across six contexts that vary by register, mode, and discourse function. Because each sentence pair differs only in lexical form, the design enables controlled testing of contextual differentiation while holding semantic content constant.

We evaluate a closed-source model (GPT-5) alongside state-of-the-art open-source Japanese-capable LLMs using three complementary tasks: sentence-level rating, pairwise comparison, and masked prediction. Together, these tasks allow us to examine both surface judgements of naturalness and token-level lexical preferences. Our study provides a reusable evaluation resource and empirical insight into how contemporary LLMs encode sociolinguistic variation in Japanese, contributing to ongoing efforts to develop culturally informed and pragmatically appropriate NLP systems. Our objective is not to determine which lexical form is correct in a given context, but to probe whether different architectures exhibit systematic and context-sensitive differentiation under controlled conditions.

2 Related Work

2.1 Loanwords in Japanese

The tripartite structure of the Japanese lexicon, *wago*, *kango*, and *gairaigo*, is well established (Shibatani, 1990; Irwin, 2011). While borrowing from Chinese has shaped the lexicon for centuries, English-derived loanwords have expanded rapidly in recent decades and now occupy a prominent role across communicative domains. Beyond filling lexical gaps, *gairaigo* frequently serve stylistic and pragmatic functions. Sociolinguistic accounts emphasise their role in indexing modernity, cosmopolitan identity, technological innovation, and global orientation (Stanlaw, 2004; Loveday, 1996; Takashi, 1990).

Loanword choice is therefore not merely semantic substitution but a socially meaningful choice. In advertising and professional discourse, English borrowings can signal Western affiliation or prestige (Takashi, 1990), while native equivalents may evoke institutional authority or tradition. Loanword choice is therefore not merely semantic substitution but a socially meaningful one. This view follows a long tradition in variationist sociolinguistics establishing that lexical and phonological alternation is socially stratified and stylistically conditioned (Labov, 1973), and that variants carry not fixed categorical meanings but a field of context-dependent social associations (Eckert, 2008). In advertising and professional discourse, English borrowings can signal Western affiliation or prestige (Takashi, 1990), while native equivalents may evoke institutional authority or tradition. Register distinctions further shape lexical selection, as formal and informal contexts systematically influence lexical, grammatical, and pragmatic choices (Liu and Allen, 2014; Dunn, 1999; Matsumoto, 1988). These findings establish lexical alternation as a context-sensitive phenomenon embedded in broader sociocultural systems.

Semantic divergence between loanwords and their English source forms has also been documented. Using distributional embeddings, Takamura et al. (2017) demonstrate measurable shifts in meaning, reinforcing the need for careful consideration of semantic equivalence when constructing interchangeable pairs. Research on English-derived words coined within Japan further highlights divergence in usage and interpretation (Hatanaka and Pannell, 2016). Such work underscores the complexity of treating loanwords and native equivalents as fully interchangeable forms.

2.2 Loanwords in Language Acquisition

Loanwords play a documented role in second language acquisition. Daulton (2008) describes English-derived vocabulary in Japanese as a "built-in lexicon" that facilitates lexical access while potentially encouraging assumptions of cross-linguistic equivalence. Classroom studies report that both learners and teachers perceive loanword-based transfer as a source of facilitation as well as confusion, particularly where form and meaning diverge (Spring, 2018). Attitudinal research further suggests ambivalence toward loanwords, balancing perceived usefulness against concerns about clarity or appropriateness (Daulton, 2011).

Empirical studies demonstrate both benefits and risks of loanword reliance. Aizawa et al. (2024) show that Japanese learners perform better on English vocabulary tests when target words correspond to familiar loanwords, while Ferries (2022) find evidence of loanword-influenced semantic transfer in learner English writing. Comprehension studies also indicate that understanding of loanwords varies depending on speaker background (Alharaki et al., 2023). Together, this literature suggests that near-equivalent loanword–native pairs may not be interpreted uniformly across audiences, and that

contextual appropriateness remains a pedagogically relevant concern.

2.3 LLM Evaluation and Sociolinguistics

Large language models have demonstrated strong performance in Japanese–English translation and related generation tasks (Yan et al., 2024; Jiao et al., 2023). Benchmarks such as the Open Japanese LLM Leaderboard evaluate translation, summarisation, and dialogue performance, indirectly reflecting stylistic competence, though without explicit focus on lexical alternation.[1]

Beyond translation quality, research has begun probing LLMs for sociolinguistic sensitivity. Controlled prompting studies show that persona and role instructions can shift output style (Salewski et al., 2023). Dialect-sensitive evaluation reveals disparities across language varieties (Deas et al., 2023; Tjuatja et al., 2024), while multilingual analyses of politeness and formality report partial but inconsistent alignment with human norms (Srinivasan and Choi, 2022). In Japanese NLP, corpora for spoken-to-written style conversion and text simplification further highlight the importance of modelling register and pragmatic variation (Ihori et al., 2020; Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018; Hatagaki et al., 2022; Nagai et al., 2024; Urakawa et al., 2024).

However, systematic evaluation of context-sensitive lexical choice between loanwords and native equivalents remains limited. Our work addresses this gap by introducing a structured evaluation for analysing contextual differentiation in Japanese lexical choice across multiple LLM architectures.

3 Data processing

3.1 Loan word extraction

To extract loanwords, we leveraged several large-scale Japanese-English datasets (JMdict Project, 2025; range3, 2023; Maruyama and Yamamoto, 2018). We filtered the lexicon to identify all entries designated as loanwords (katakana written terms) and selected those with a corresponding native Japanese synonym. This procedure provided an initial list of pairs. This list was further supplemented with manually selected pairs from semantic domains in which loanwords are especially prevalent (e.g., colour terminology), ensuring broader coverage across frequently occurring lexical categories. The resulting list was then curated to retain only those pairs where the terms were commonly interchangeable in modern usage. In total, we extracted 221 candidate loanwords.

3.2 Sentence Generation

Using the curated list of loanword–native pairs, we constructed a sentence-level evaluation dataset. To ensure systematic and context-controlled generation, we employed the generative model Gemini 2.5 Pro (Comanici et al., 2025) to produce paired sentences for each lexical item. The aim was to embed each pair within communicative settings that vary along dimensions of register, mode, and discourse function, enabling controlled testing of contextual sensitivity. Prompt provided in Appendix B in Tab 5.

For each lexical pair, we generated sentences across six predefined communicative contexts:

- **Formal Conversation:** A short, polite dialogue typical of a business or service interaction.
- **Formal Written:** A sentence resembling a report, academic paper, or official correspondence.
- **Formal Explanation:** A definition or technical description, as found in a textbook or manual.
- **Informal Conversation:** A casual exchange between friends or peers.
- **Informal Written:** A sentence similar to a personal message, email, or social media post.
- **Informal Explanation:** A casual explanation directed at a peer.

For each context, the model was prompted to generate two sentences that were semantically equivalent and differed only in lexical choice, specifically the use of the loanword versus its native counterpart. Because some loanwords exhibit polysemy, prompts explicitly constrained generation to senses in which both forms were contextually interchangeable. This ensured that lexical form remained the sole manipulated variable, isolating contextual preference rather than semantic divergence.

3.3 Automatic Quality Check

We implemented an automatic quality check to filter the generated dataset for semantic consistency. The core of this check was a back-translation workflow designed to detect potential meaning shifts between the loanword and native sentence variants.

[1] https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard

Japanese	English
これは成功するチャンスです。	This is a **chance** to succeed.
これは成功する機会です。	This is an **opportunity** to succeed.

Table 1: Example illustrating lexical pair: “チャンス” (chansu) vs. “機会” (kikai).

For each Japanese sentence pair, both versions were translated into English using the DeepL API[2] and the Google Translate API,[3] producing corresponding English sentences. To quantify semantic similarity, we generated sentence embeddings for each translation using a pre-trained multilingual Sentence-BERT model (Reimers and Gurevych, 2019).[4] Cosine similarity was then computed between the embedding vectors.

Sentence pairs with a similarity score below a threshold of 0.95 were automatically flagged for review. Terms for which the majority of sentence pairs were flagged were removed from the dataset. For terms with only one or two flagged instances, new sentences were generated.

The use of round-trip translation as a proxy for semantic equivalence follows established practice in machine translation evaluation, where back-translation and semantic comparison are used to detect meaning divergence and verify consistency between parallel sentence pairs (Jia et al., 2025; Edunov et al., 2020; Federmann, 2012).

3.4 Manual Quality Check

The quality and consistency of the generated dataset were ensured through manual verification by a fluent bilingual (English–Japanese) evaluator. The evaluator confirmed the semantic integrity and contextual interchangeability of each sentence pair. Specifically, both the loanword and native term had to convey the same propositional meaning and function as valid substitutes within the specified context (e.g., Formal Written). Grammatical well-formedness under substitution was also verified.

Table 1 illustrates a case in which lexical alternation may introduce subtle shifts in connotation without altering overall intent. Although *chansu* (チャンス) and *kikai* (機会) both refer to the possibility of doing something, chansu is often rendered as “chance,” carrying a slightly casual or motivational tone, whereas kikai is more frequently translated as “opportunity,” particularly in formal or institutional contexts. Substituting one for the other does not substantially change the underlying meaning of the sentence, but it may affect perceived register or formality. Pairs exhibiting such shifts in domain specificity, connotation, or pragmatic scope were excluded from the dataset. This filtering criterion was applied consistently across all candidates to ensure that retained pairs were genuinely interchangeable across the six communicative contexts. After automatic and manual quality checks, 113 lexical pairs remained. Each pair was embedded across six contexts with two variants per context, yielding 12 sentences per pair and a total of **1,356** sentences in the final dataset.

4 Experimental Setup

4.1 Task Definition

We evaluate lexical selection using three complementary tasks designed to probe contextual sensitivity at both sentence and token levels.

In the **Rating Task**, the model is presented with a single sentence containing either the loanword or its native equivalent, together with the specified communicative context. The model assigns a naturalness score on a five-point scale. These ratings allow comparison of perceived acceptability across lexical forms and contexts.

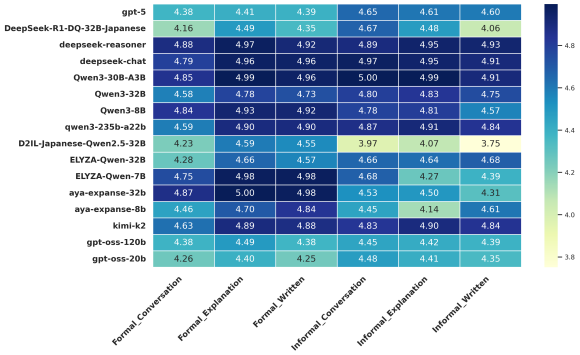
In the **Comparison Task**, the model is given two sentences that are identical except for the target word and must select the more natural option within the given context. Each pair is presented twice with reversed order. The **Self** score in this task measures order consistency, calculated as the proportion of instances in which a model selects the same lexical option regardless of presentation order. Pairwise agreement across models is used to assess cross-model convergence in lexical preference.

In the **Masked Prediction Task**, the target word is removed and the model is asked to generate the most appropriate lexical item. We record whether the model produces the loanword, the native equivalent, or an alternative form. Two variants are implemented. In the **WITH** condition, the model selects between the original pairs. In the **OPEN** condition, no restriction is imposed and the model freely generates a lexical item. The **Self** score in this task measures agreement between a model’s WITH and OPEN predictions for the same item.

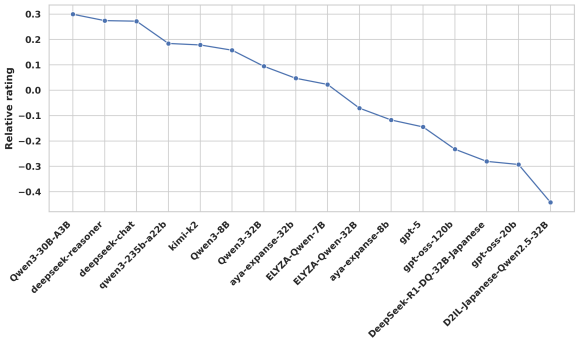
[2] https://www.deepl.com/en/pro-api

[3] https://cloud.google.com/translate/docs/reference/rest

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2



(a) Average ratings heatmap.



(b) Model rating sensitivity.

Figure 1: Results of the Rating Task. (a) Average naturalness ratings (1–5) across models and contexts. (b) Context sensitivity of models, shown as relative differences between model scores and the overall average.

Importantly, our evaluation does not treat either lexical form as inherently correct within a given context, nor does it assume a fixed normative standard of speaker preference. Rather than measuring alignment with human speaker norms, our analysis focuses on relative architectural divergence across models under controlled contextual conditions. The goal is to examine how different LLMs distribute lexical choices given identical inputs, not to determine which model best reflects contemporary usage.

4.2 Models

We evaluate a broad set of Japanese-capable LLMs. The closed-source group is represented by `GPT-5` (Singh et al., 2025), while the open-source group spans several model families: `GPT-OSS` (20B, 120B)(OpenAI, 2025), `DeepSeek` (Reasoner, Chat) (DeepSeek-AI, 2024), `CyberAgent` (DeepSeek-R1 Distill Qwen-32B Japanese) (Ishigami, 2025), `Deep Analysis Research` (Japanese Qwen2.5-32B), the `ELYZA Shortcut series` (7B, 32B) (Hirakawa et al., 2025), the `Qwen-3 family` (8B, 32B, 30B-A3B, 235B-A22B) (Team, 2025), `Kimi-K2` (Team et al., 2025), and `Cohere Aya-Expanse` (8B, 32B) (Dang et al., 2024). All prompts and model specifications are provided in Appendix A and B.

5 Results

5.1 Rating Task

Although both loanword and native equivalents are valid, one may be more appropriate in context. The Rating Task evaluates whether the constructed pairs are perceived as interchangeable by measuring overall naturalness without separating scores by lexical type. Across all models, ratings for sentences containing either form clustered toward the upper end of the scale (see Figure 1a). This concentration suggests that the majority of generated pairs were judged natural regardless of whether the loanword or native variant was used, supporting the semantic equivalence of the dataset.

Within this generally compressed range, variation is more strongly attributable to model calibration than to lexical form. Stability differs by architecture rather than uniformly by size. DeepSeek-Reasoner exhibits one of the flattest rating profiles across communicative settings, with minimal separation between formal and informal contexts, and DeepSeek-Chat shows similarly limited spread. GPT-5, by contrast, demonstrates a systematic uplift in informal settings relative to formal ones, indicating a consistent context effect rather than complete uniformity. DeepSeek-R1-DQ-32B-Japanese displays the largest within-model variation, largely driven by a lower score in Informal Written. Family-level tendencies are also visible: Qwen and DeepSeek models generally assign higher average ratings overall, whereas GPT-OSS variants apply comparatively stricter evaluations, as illustrated in Figure 1b.

These differences primarily reflect evaluative calibration rather than strong lexical discrimination. More generous systems may present a broad range of lexical choices as equally acceptable, potentially obscuring finer register distinctions. Conversely, stricter systems may assign lower scores even when both variants are contextually legitimate. The Rating Task confirms that both variants are generally accepted as natural within their contexts. This ensures that subsequent tasks examine differences in preference rather than problems of semantic mismatch.

Comparative research in Japanese sociolinguistics shows that lexical preferences vary systematically across contexts. Native terms are generally favoured in formal settings, reflecting expectations of careful or official language use (Hashimoto, 2019). In informal contexts, distinctions are less stable, and conversational settings tend to allow greater variability (Stanlaw, 2004; Loveday, 1996). Divergence between lexical alternatives is often greater in interactionally sensitive contexts and lower in informational or technical discourse (Stanlaw, 2004; Loveday, 1996). Sensitivity to register and context is therefore central to sociolinguistic norms of lexical selection.

The relative generosity of Qwen and DeepSeek models may reflect differences in training objectives and instruction tuning. Systems optimised for conversational helpfulness or safety alignment may be less inclined to assign low ratings in the absence of clear grammatical errors. Training data composition may also play a role, as exposure to web or media corpora, where katakana loanwords are frequent, could increase tolerance towards lexical variation. Sentence length likewise affects ratings (Appendix C). Overall, rating behaviour appears to reflect calibration and optimisation choices in addition to sociolinguistic sensitivity, underscoring the importance of recognising model-specific tendencies when interpreting LLM feedback.

5.2 Comparison Task

When comparing sentence pairs, models exhibited systematic but non-convergent preferences. Agreement rates were consistently above chance yet moderate overall (see Table 2). Larger models tended to align more closely with one another, with GPT-5 showing the highest agreement with GPT-OSS-120B and DeepSeek-Reasoner. In contrast, smaller Aya, Kimi, and ELYZA variants demonstrated lower cross-system alignment.

Order bias was evaluated by reversing sentence presentation (Pezeshkpour and Hruschka, 2024). Most models maintained high internal consistency under this manipulation, particularly larger systems, indicating that preferences were stable and unlikely to arise from superficial prompt artefacts. Model disagreement therefore appears to reflect genuine differences in lexical judgement rather than task instability.

Lexical preferences further revealed a contextual divide (Figure 2). Native equivalents were more frequently selected in formal registers, whereas

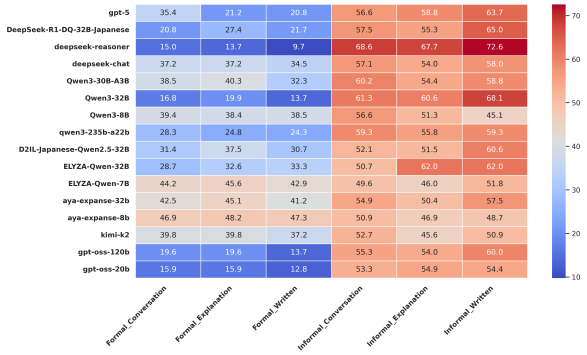


Figure 2: Heatmap of loanword selection percentages across six communicative contexts. Values indicate how often models chose the loanword sentence. Darker/red = higher loanword selection; lighter/blue = native selection

loanwords appeared more often in informal settings. DeepSeek and GPT-OSS variants demonstrated clearer contextual sensitivity, adjusting preferences across registers. Aya variants showed weaker differentiation and tended to favour loanwords more uniformly. We do not establish a human normative baseline in this study; our analysis therefore describes divergence in model-internal lexical distributions rather than verified alignment with contemporary speaker behaviour.

The contrast with the Rating Task is notable. While scalar ratings suggested broad acceptability of both forms, direct comparison exposed sharper architectural divergence. By requiring an explicit binary choice, the Comparison Task reveals lexical preferences that remain hidden in gradient evaluations.

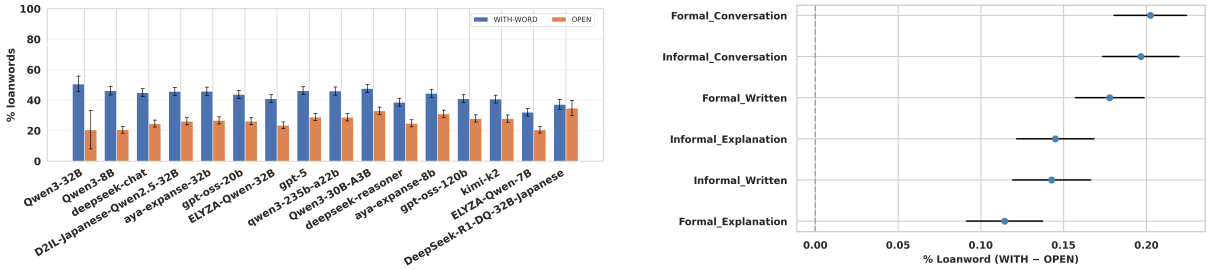
5.3 Masked Word Prediction Task

Within-model "Self" scores measured consistency between the WITH condition, in which the original loanword–native pair was provided, and the OPEN condition, which allowed unconstrained generation. Most systems demonstrated moderate to high internal stability, particularly ELYZA, Kimi and GPT variants. However, internal consistency did not necessarily translate into cross-model agreement. ELYZA models were stable within themselves yet systematically distinct from other systems, whereas Qwen models showed lower internal stability but aligned more closely with peers.

Pairwise comparisons under both WITH and OPEN exceeded chance levels but varied across model families as shown in Table 3. DeepSeek-Chat, Aya-Expanse-32B and Qwen3-32B exhibited relatively stronger agreement with other systems,

Model	Self (%)	Pairwise agreement (%)				
		Avg	Min	Model	Max	Model
gpt-5	91.7	68.5	56.8	aya-expanse-8b	78.5	gpt-oss-120b
gpt-oss-120b	85.3	66.6	55.5	aya-expanse-8b	78.5	gpt-5
deepseek-reasoner	85.1	64.6	57.2	kimi-k2	75.8	gpt-5
gpt-oss-20b	83.3	65.3	55.1	aya-expanse-8b	75.4	gpt-5
Qwen3-32B	82.3	64.4	54.9	aya-expanse-8b	73.0	gpt-5
DeepSeek-R1-DQ-32B-Japanese	74.0	63.0	56.4	ELYZA-Qwen-7B	69.0	deepseek-reasoner
ELYZA-Qwen-32B	68.0	61.5	55.1	ELYZA-Qwen-7B	65.0	gpt-5
Qwen3-8B	63.0	60.0	54.1	aya-expanse-8b	64.6	gpt-5
aya-expanse-32b	62.4	60.4	55.0	aya-expanse-8b	65.9	gpt-5
qwen3-235b-a22b	61.8	61.3	53.8	aya-expanse-8b	68.0	gpt-5
Qwen3-30B-A3B	59.1	59.3	53.6	aya-expanse-8b	65.2	gpt-5
D2IL-Japanese-Qwen2.5-32B	59.1	60.1	53.6	aya-expanse-8b	65.6	gpt-5
deepseek-chat	55.2	60.5	53.9	aya-expanse-8b	65.5	gpt-5
ELYZA-Qwen-7B	46.5	57.2	53.5	kimi-k2	60.3	gpt-5
kimi-k2	33.8	56.6	52.2	aya-expanse-8b	58.8	gpt-5
aya-expanse-8b	32.9	54.6	52.2	kimi-k2	56.8	gpt-5

Table 2: Pairwise agreement. "Self" is to check order bias. Avg/Max/Min computed excluding self.



(a) Loanword selection rates under WITH vs. OPEN.

(b) Contextual differences between WITH and OPEN.

Figure 3: Loanword usage patterns in the Masked Prediction Task. (a) shows aggregated loanword proportions across models under WITH vs. OPEN with order from left to right showing the difference between the two tasks; (b) shows context-specific gaps with 95% confidence intervals.

while ELYZA-Qwen-7B and DeepSeek-R1-DQ-32B showed weaker alignment. Loanword selection patterns clarify these dynamics (see Figure 3). Under WITH, loanword usage approached balance between alternatives. Under OPEN, loanword rates decreased and cross-model divergence increased. Constraining the candidate set therefore promotes convergence, whereas unconstrained generation exposes underlying distributional tendencies. Task design thus meaningfully shapes lexical outcomes.

Taken together, masked prediction shows that models converge at the level of broad stylistic orientation but diverge at the level of specific lexical realisation. Structured prompting encourages agreement, whereas open generation surfaces architectural differences in lexical priors.

5.4 Discussion

Our results show that sentence-level ratings conceal substantial variation in lexical preference. Across contexts, models consistently assigned high naturalness scores to both loanword and native alternatives, creating the impression of broad acceptability. However, comparison and masked prediction tasks revealed clearer divergence: models differed in loanword frequency and in the degree to which they tracked contextual cues. In several cases, architectural divergence exceeded shifts across communicative contexts. Rating-based evaluation therefore appears relatively flat, whereas token-level probing exposes finer lexical tendencies. This pattern aligns with findings that generative models can exhibit socially patterned behaviour even when surface-level evaluations suggest neutrality (Hu et al., 2025).

Agreement patterns further clarify this distinction. Models often converged on broad stylistic orientation while diverging in specific lexical realisations. Masked OPEN prediction tended to favour native forms, whereas constrained WITH prediction produced more balanced distributions. Pairwise comparisons also revealed clearer register-based shifts than scalar ratings (Figure 4). These findings support the view that language models encode distributions of socially meaningful styles and

Model	Self (%)	WITH (%)					OPEN (%)				
		Avg	Min	Model	Max	Model	Avg	Min	Model	Max	Model
ELYZA-Qwen-7B	73.7	74.3	71.7	gpt-5	78.1	ELYZA-Qwen-32B	75.8	71.5	DeepSeek-R1-DQ-32B-Japanese	80.0	deepseek-chat
kimi-k2	72.5	79.5	76.2	ELYZA-Qwen-7B	83.6	deepseek-chat	76.5	68.6	DeepSeek-R1-DQ-32B-Japanese	87.2	Qwen3-32B
gpt-oss-120b	72.3	79.2	74.4	ELYZA-Qwen-7B	83.0	deepseek-chat	75.7	67.5	DeepSeek-R1-DQ-32B-Japanese	81.0	deepseek-chat
gpt-5	71.8	78.9	71.7	ELYZA-Qwen-7B	86.2	Qwen3-32B	73.5	64.7	DeepSeek-R1-DQ-32B-Japanese	77.9	gpt-oss-120b
deepseek-reasoner	71.3	78.3	75.6	ELYZA-Qwen-7B	83.4	deepseek-chat	76.3	69.9	DeepSeek-R1-DQ-32B-Japanese	85.4	Qwen3-32B
ELYZA-Qwen-32B	70.4	80.2	76.7	gpt-oss-20b	82.1	D2IL-Japanese-Qwen2.5-32B	78.0	72.1	DeepSeek-R1-DQ-32B-Japanese	84.6	Qwen3-32B
deepseek-chat	69.8	81.4	74.5	ELYZA-Qwen-7B	85.4	aya-expanse-32b	78.0	69.7	DeepSeek-R1-DQ-32B-Japanese	81.7	ELYZA-Qwen-32B
qwen3-235b-a22b	69.6	79.3	73.1	ELYZA-Qwen-7B	83.3	aya-expanse-32b	75.4	68.8	DeepSeek-R1-DQ-32B-Japanese	79.0	deepseek-chat
aya-expanse-8b	67.8	79.1	74.2	ELYZA-Qwen-7B	83.9	aya-expanse-32b	72.9	67.2	DeepSeek-R1-DQ-32B-Japanese	79.2	aya-expanse-32b
Qwen3-30B-A3B	66.9	78.5	74.0	ELYZA-Qwen-7B	80.6	ELYZA-Qwen-32B	72.3	66.2	DeepSeek-R1-DQ-32B-Japanese	76.7	Qwen3-32B
aya-expanse-32b	66.7	81.5	75.7	ELYZA-Qwen-7B	85.4	deepseek-chat	77.3	70.4	DeepSeek-R1-DQ-32B-Japanese	84.2	Qwen3-32B
gpt-oss-20b	66.6	77.2	72.1	ELYZA-Qwen-7B	81.8	Qwen3-32B	73.5	68.1	DeepSeek-R1-DQ-32B-Japanese	81.4	Qwen3-32B
DeepSeek-R1-DQ-32B-Japanese	66.3	79.6	77.5	ELYZA-Qwen-7B	82.0	ELYZA-Qwen-32B	69.9	64.7	gpt-5	85.7	Qwen3-32B
D2IL-Japanese-Qwen2.5-32B	66.1	76.8	72.2	ELYZA-Qwen-7B	82.1	ELYZA-Qwen-32B	74.4	67.7	DeepSeek-R1-DQ-32B-Japanese	81.4	Qwen3-32B
Qwen3-8B	61.6	77.4	72.6	ELYZA-Qwen-7B	80.3	Qwen3-30B-A3B	74.6	70.3	DeepSeek-R1-DQ-32B-Japanese	79.1	deepseek-chat
Qwen3-32B	60.0	80.7	73.3	ELYZA-Qwen-7B	86.2	gpt-5	79.5	72.5	gpt-oss-120b	87.2	kimi-k2

Table 3: Pairwise agreement summary by condition (WITH / OPEN). "Self" is WITH vs. OPEN self-consistency. All values shown as percentages. Avg/Max/Min computed excluding self.

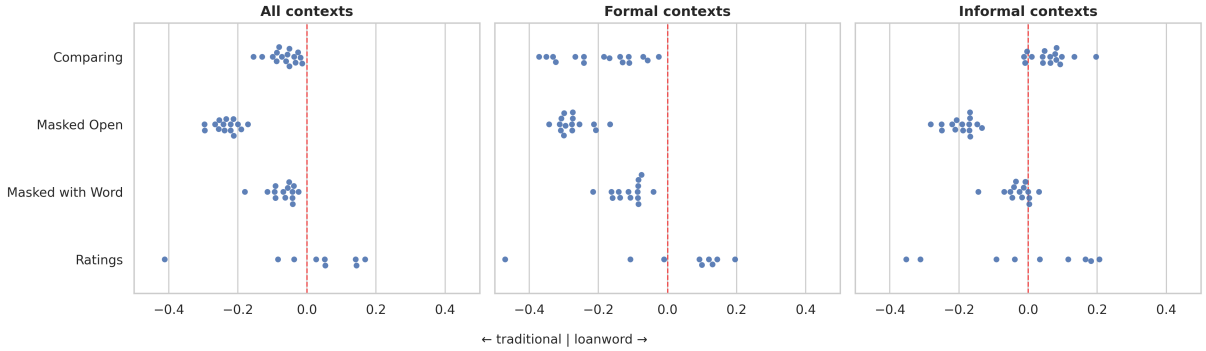


Figure 4: Bias scores across tasks (ratings, comparisons and masked predictions) in formal and informal contexts. Points represent model averages; values to the left indicate preference for native forms, and values to the right indicate preference for loanwords.

registers alongside grammatical structure (Grieve et al., 2025).

At the same time, register alone does not determine loanword usage. English-derived forms in Japanese frequently index modernity, global orientation, technical precision, or euphemistic framing independently of formal–informal contrasts (Stanlaw, 2004; Loveday, 1996; Takashi, 1990). In explanatory or technical contexts, katakana forms may signal domain alignment rather than informality. While our evaluation operationalises register within structured settings, the observed variation likely reflects interaction between contextual cues and broader sociocultural meanings. The results should therefore be interpreted as evidence of register-sensitive differentiation within this design rather than as a comprehensive account of the indexical functions associated with loanword usage.

These findings also have implications for LLM-assisted language learning. Scalar judgements may present alternatives as equally natural, obscuring contextual differentiation, whereas comparison and generative tasks reveal architecture-dependent lexical tendencies that vary with prompt framing. Effective deployment of LLMs in educational contexts therefore requires sociolinguistically informed evaluation and careful calibration (Nguyen, 2025). This dynamic is reminiscent of how language attitudes are socially conditioned: perceptions of appropriateness and prestige are shaped by social experience rather than being fixed properties of forms (Garrett, 2010; Eckert, 2008). In contemporary Japanese discourse, loanwords often function as markers of modernity and professional identity (Takashi, 1990). If training data disproportionately reflect media-heavy corpora, similar distributional pressures may influence model outputs, consistent with research on implicit bias formation in humans and language models (Greenwald and Banaji, 1995; Caliskan et al., 2017). While prompting methods can alleviate some of these tendencies by constraining output space or explicitly foregrounding register, they do not eliminate underlying distributional biases inherited from training data and model optimisation.

Although designed for controlled evaluation, the dataset makes explicit how lexical alternation interacts with communicative context by embedding interchangeable pairs across six register conditions. This structure isolates register-sensitive variation rather than inferring it from heterogeneous corpora.

More broadly, the results underscore that surface-level naturalness does not guarantee contextual appropriateness, highlighting the importance of task design in probing socially conditioned lexical choice in LLMs.

6 Conclusion

Our findings demonstrate that LLMs do not encode a stable contextual rule for loanword versus native lexical selection, despite producing generally fluent outputs. While both forms are often judged acceptable at the sentence level, cross-architectural differences reveal uneven sensitivity to socially conditioned register cues. This highlights a broader limitation in socially grounded language modeling: fluency does not guarantee consistent representation of pragmatic variation. For learner-facing applications, this means outputs should be interpreted cautiously and framed appropriately. More broadly, our dataset provides a controlled benchmark for evaluating context-sensitive lexical generation and offers insight into how contemporary LLMs model socially meaningful linguistic alternation.

Limitations

Our study isolates lexical choice by constructing semantically equivalent sentence pairs. While this provides experimental control, it overlooks document context, which can strongly influence lexical decisions in communication. The process of selecting lexical pairs and generating sentences with LLM assistance, followed by back-translation using commercial MT systems, may introduce uneven coverage of semantic domains and artifacts from machine translation. Finally, our focus on English-derived loanwords in general-domain contexts limits the scope of our findings, they may not generalise to loanwords from highly specialised terminology.

Japanese presents a particularly demanding test case for lexical modelling because sociolinguistic contrasts are partially encoded orthographically (e.g., katakana vs. kanji), historically layered (wago/kango/gairaigo), and pragmatically dependent on discourse setting. A model that succeeds in distinguishing these layers demonstrates sensitivity not only to lexical frequency but to socially structured linguistic choice. This makes Japanese a high-resolution benchmark for stylistic control in LLMs more broadly.

Acknowledgements

Joseph James was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, and 1 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Sura Alharaki, Muhammad Alif Redzuan Abdullah, and Syed Nurulakla Bin Syed Abdullah. 2023. Comprehension of english loanwords in japanese by japanese and english speakers. *World*, 13(5).

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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.

Frank E Daulton. 2008. *Japan's built-in lexicon of English-based loanwords*, volume 26. Multilingual Matters.

Frank E Daulton. 2011. On the origins of gairaigo bias: English learners' attitudes towards english-based loanwords in japan. *The Language Teacher*, 35:7.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Cynthia Dickel Dunn. 1999. Coming of age in japan: Language ideology and the acquisition of formal speech registers. In *Language and ideology: selected papers from the Sixth International Pragmatics Conference*, volume 1, pages 89–97. International Pragmatics Association Antwerp.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *Prague Bull. Math. Linguistics*, 98:25–36.

Jonathan Ferries. 2022. A corpus analysis of loanword effects on second language production. *Englishes in Practice*, 5(1):107–132.

Peter Garrett. 2010. *Attitudes to language*. Cambridge University Press.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Daiki Hashimoto. 2019. Sociolinguistic effects on loanword phonology: Topic in speech and cultural image. *Laboratory Phonology*, 10(1).

Koki Hatagaki, Tomoyuki Kajiwara, and Takashi Ninomiya. 2022. Parallel corpus filtering for japanese text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 12–18.

Mariko Hatanaka and Justin Pannell. 2016. English loanwords and made-in-japan english in japanese. *Hawaii Pacific University TESOL Working Paper Series*, 14:14–29.

Masato Hirakawa, Tomoaki Nakamura, Akira Sasaki, Daisuke Oba, and Shoetsu Sato. 2025. elyza/elyza-thinking-1.0-qwen-32b.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Mana Ihori, Akihiko Takashima, and Ryo Masumura. 2020. Parallel corpus for japanese spoken-to-written style conversion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6346–6353.

Mark Irwin. 2011. Loanwords in japanese.

Ryosuke Ishigami. 2025. Deepseek-r1-distill-qwen-32b-japanese.

Yepai Jia, Yatu Ji, Xiang Xue, Lei Shi, Qing-Dao-Er-Ji Ren, Nier Wu, Na Liu, Chen Zhao, and Fu Liu. 2025. A semantic uncertainty sampling strategy for back-translation in low-resources neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 528–538, Vienna, Austria. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

JMdict Project. 2025. JMdict Japanese–English Dictionary (Yomitan distribution). `https://github.com/yomidevs/jmdict-yomitan`. Accessed 2025-02-16.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Xiangdong Liu and Todd James Allen. 2014. A study of linguistic politeness in japanese. *Open Journal of Modern Linguistics*, 4(05):651–663.

Leo J Loveday. 1996. *Language contact in Japan: A sociolinguistic history*. Clarendon Press.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of pragmatics*, 12(4):403–426.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

range3. 2023. Japanese Wikipedia Dump (2023-01-01 version). https://huggingface.co/datasets/range3/wikipedia-ja-20230101. Accessed 2025-02-16.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Mark Spring. 2018. Unconscious gairaigo bias in efl: A case study of japanese teachers of english. *Shinshu University Journal of Arts and Sciences*, 12:166–181.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Stanlaw. 2004. *Japanese English: Language and culture contact*, volume 1. Hong Kong University Press.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Kyoko Takashi. 1990. A sociolinguistic analysis of english borrowings in japanese advertising texts. *World Englishes*, 9(3):327–341.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Takako Tomoda. 1999. The impact of loan-words on modern japanese. In *Japan Forum*, volume 11, pages 231–253. Taylor & Francis.

Toru Urakawa, Yuya Taguchi, Takuro Niitsuma, and Hideaki Tamori. 2024. A Japanese news simplification corpus with faithfulness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 659–665, Torino, Italia. ELRA and ICCL.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

A Models

Open-source models were either run locally on an H100 80GB GPU or accessed via LiteLLM[5], while closed-source models were accessed through their official APIs. All evaluations were conducted under default settings. Full model list is provided in Table 4.

[5] https://docs.litellm.ai/docs/project

Family	Size	Model ID
OpenAI (Closed)	–	gpt-5-2025-08-07
GPT-OSS	120B	openai/gpt-oss-120b
	20B	openai/gpt-oss-20b
DeepSeek	685B	deepseek/deepseek-reasoner
	685B	deepseek/deepseek-chat
Cohere Aya-Expanse	32B	CohereLabs/aya-expanse-32b
	8B	CohereLabs/aya-expanse-8b
ELYZA Shortcut	7B	elyza/ELYZA-Shortcut-1.0-Qwen-7B
	32B	elyza/ELYZA-Shortcut-1.0-Qwen-32B
Qwen-3 Family	30B	Qwen/Qwen3-30B-A3B-Instruct-2507
	235B	qwen/qwen3-235b-a22b-instruct-2507
	4B	Qwen/Qwen3-4B-Thinking-2507
	32B	Qwen/Qwen3-32B
	8B	Qwen/Qwen3-8B
Kimi	1T	kimi-k2-instruct
Deep Analysis Research	32B	deep-analysis-research/D2IL-Japanese-Qwen2.5-32B-Instruct-v0.1
CyberAgent	32B	cyberagent/DeepSeek-R1-Distill-Qwen-32B-Japanese

Table 4: Models evaluated.

B Prompts

You are a data generation bot for a linguistics research project. Your task is to take a Japanese loanword and its traditional counterpart and generate a complete dataset entry.

Your entire output must strictly follow the format specified below. Do not include any introductory text, explanations, headers, or anything else outside of this format.

1. First Line: Provide the single English word that is the common translation for the input pair.
2. Subsequent Six Lines: For each of the six "sentence topics" below, generate one complete data row.
   - Crucially, you must generate a pair of sentences for each topic: one using the loanword and one using the traditional word.
   - Ensure the two sentences are semantically identical and the words are as interchangeable as possible within that context. The only difference should be the target words.
   - Provide a single, shared English translation for the pair.

The six sentence topics are:
- Formal_Conversation
- Formal_Written
- Formal_Explanation
- Informal_Conversation
- Informal_Written
- Informal_Explanation

Output Format:

Your output must be structured in exactly 7 lines: one line for the translated word, and six lines for the data, formatted as follows. Use | as the delimiter.

[Translated Word]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]

Input Word Pair:
* Loanword: {loanword}
* Traditional Word: {traditional_word}

Table 5: System prompt for dataset generation using Gemini 2.5 Pro.

You are an AI Japanese language model. Given a single Japanese sentence and the **sentence type** specified by the user (e.g., casual statement, polite request, formal announcement), assess how well the sentence:

1. follows Japanese grammar,
2. sounds natural to native speakers, and
3. suits the intended type in terms of register and word choice nuance.

Respond in exactly two sections:

1. Overall Rating (1-5): Place the single integer score (1-5) inside <score>(1-5)</score> tags.
2. Brief Analysis (1 sentence)
 • State the main issue (grammar, unnatural wording, register).
 • Mention any words you would replace (if any).
 • If the rating is 5, simply note that the sentence is well-formed and appropriate.

Keep the analysis concise; do not add extra sections or explanations.

Table 6: System prompt used for the rating task.

You are an AI Japanese language model. You will be given a context and two Japanese sentences, labeled 'a' and 'b'.

Your task is to determine which sentence sounds more natural and is more appropriate for the given context.

Respond with your choice, 'a' or 'b', inside <choice> tags.

For example:

<choice>[choice]</choice>.

Do not provide any other text, explanation, or punctuation.

Table 7: System prompt used for the comparison task.

You are an AI Japanese language model. Your task is to predict the most likely word to fill the blank space marked with [MASK] in the provided sentence. Based on the context, provide the most probable word that could complete the sentence. Respond ONLY with the word. The word must be enclosed in <option> tags. Do not add any other text, explanations, or numbering. Example format: <option>[option]</option>

Table 8: System prompt used for the open masked prediction task.

You are an AI Japanese language model. Your task is to find the single best Japanese word to fill the [MASK] in a sentence. You will be given the sentence and the target English word that the mask represents. Based on the context of the sentence and the meaning of the English word, provide the single most probable Japanese word. Respond ONLY with the Japanese word, enclosed in <option> tags. Do not add any other text or explanations. Example format: <option>[option]</option>

Table 9: System prompt used for the masked prediction task.

C Rating score based on sentence length

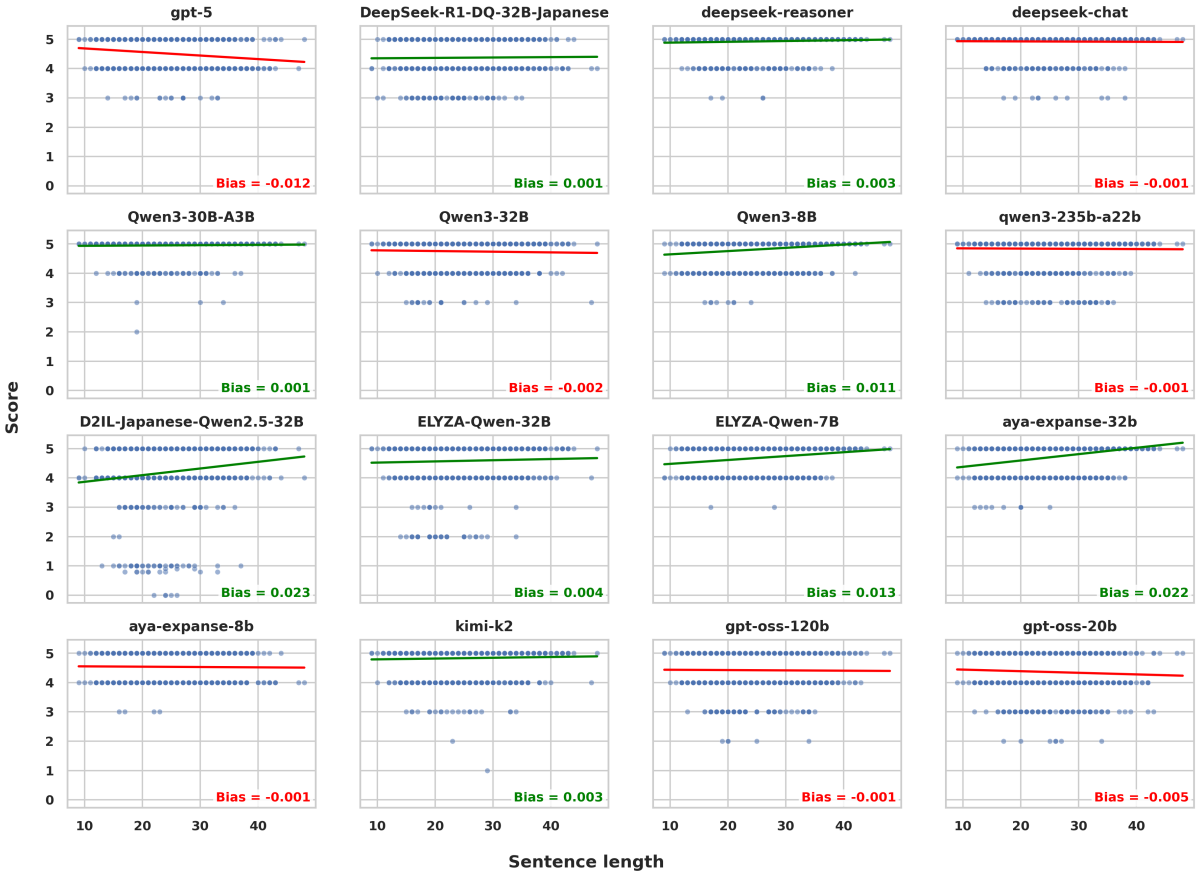


Figure 5: Scores by sentence length (number of characters) across models, where the slope indicates length bias.