Doppelganger-JC: Benchmarking the LLMs' Understanding of Cross-Lingual Homographs between Japanese and Chinese

Yuka Kitamura [1,3], **Jiahao Huang** [1], **Akiko Aizawa** [1,2,3]
[1]the University of Tokyo, [2]NII, [3]NII LLMC
ykitamura@nii.ac.jp
jiahao-huang@g.ecc.u-tokyo.ac.jp
aizawa@nii.ac.jp

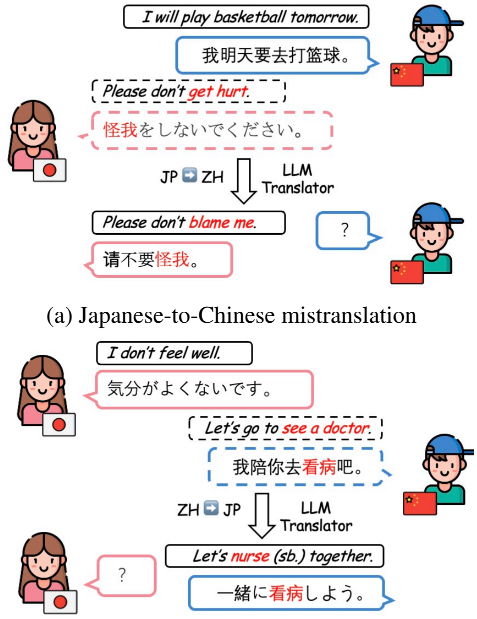
Abstract

The recent development of LLMs is remarkable, but they still struggle to handle cross-lingual homographs effectively. This research focuses on the cross-lingual homographs between Japanese and Chinese—the spellings of the words are the same, but their meanings differ entirely between the two languages. We introduce a new benchmark dataset named Doppelganger-JC to evaluate the ability of LLMs to handle them correctly. We provide three kinds of tasks for evaluation: word meaning tasks, word meaning in context tasks, and translation tasks. Through the evaluation, we found that LLMs' performance in understanding and using homographs is significantly inferior to that of humans. We pointed out the significant issue of homograph shortcut, which means that the model tends to preferentially interpret the cross-lingual homographs in its easy-to-understand language. We investigate the potential cause of this homograph shortcut from a linguistic perspective and pose that it is difficult for LLMs to recognize a word as a cross-lingual homograph, especially when it shares the same part-of-speech (POS) in both languages. The data and code is publicly available here: https://github.com/0017-alt/Doppelganger-JC.git.

1 Introduction

In recent years, the rapid development of large language models (LLMs) has led to their widespread application in various tasks. In most cases, LLMs are English-centric because there are rich data sources in English, but multilingualism is essential now (Kargaran et al., 2024).

Recent research (Tanwar et al., 2025) points out that LLMs cannot deal well with homographs, especially under bilingual settings. We have also observed similar issues when using LLMs for translation tasks in practice. Taking Chinese and Japanese as examples, these two languages have historically influenced each other, resulting in a significant number of shared words across both languages. Some of these shared words retain similar meanings, but others have undergone semantic shifts and now possess entirely different meanings in the two languages, which are known as cross-lingual homographs. Statistically, among the top 20,000 most frequently used words in Japanese, 50% originate from Chinese words, 29% are homophones, and 6% are cross-lingual homographs (松下達彦 et al., 2020). It is very important to deal properly with them, but we have found that LLMs sometimes misuse cross-lingual homographs during translation tasks, directly transferring words from the source language to the target language without proper translation, even when their meanings differ completely.



(a) Japanese-to-Chinese mistranslation

(b) Chinese-to-Japanese mistranslation

Figure 1: Examples of mistranslations in both directions

Figure 1 illustrates two examples of the misuse of cross-lingual homographs. This tendency

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1779–1794
December 20-24, 2025 ©2025 Association for Computational Linguistics

of LLMs can lead to misunderstandings in cross-linguistic communication and thus warrants attention and resolution. However, to the best of our knowledge, we couldn't identify any available datasets for experimental purposes. Therefore, in this paper, we (1) construct a dataset **Doppelganger-JC** for benchmarking the understanding of cross-lingual homographs; (2) conduct an in-depth analysis of model performance and underlying causes of this phenomenon.

Doppelganger-JC is a benchmark designed to evaluate and analyze LLMs' understanding of cross-lingual homographs between Japanese and Chinese[1]. Therefore, the benchmark includes both Japanese and Chinese, with three types of tasks for each language: word meaning, word meaning in context, and sentence translation. These tasks are intended to assess whether LLMs can correctly interpret the meanings of homographs in both languages. We primarily employed multiple-choice question and answer (MCQA) tasks and carefully designed distraction options, while also supplementing our analysis with open-ended tasks. We obtained LLMs' predicted answers by selecting the option with the lowest perplexity. Further details on the dataset generation process are provided in § 3.

We evaluated the performance of two Japanese models, two Chinese models, and three other models on our dataset to evaluate the effect of languages in the training corpus (as shown in Table 1).

Model	Pre-training Corpus Language
llm-jp-3-7.2b-instruct3 (LLM-jp et al., 2024)	English (950B), Japanese (592B), Korean (0.3B), Chinese (0.8B)[2]
Llama-3-ELYZA-JP-8B (Hirakawa et al., 2024)	Japanese (further pre-training), primarily English, and 5% of other languages over 30 (Llama-3-8B-Instruct[3])
Qwen2.5-7B-Instruct-1M (Yang et al., 2025a; Team, 2025)	30 languages (such as English, Chinese, Spanish, French, German, Arabic, Russian, Korean, Japanese, Thai, and Vietnamese) (Yang et al., 2024)
Baichuan2-7B-Base (Yang et al., 2023)	Although not mentioned, it is stated that multilingualism has been acquired.
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	primarily English, and 8% of other languages over 176
gemma-7b (Team et al., 2024)	primarily English
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Although not mentioned, it supports English and Hinglish[4]

Table 1: Models we use in this research.

We found that the models' comprehension of sentences containing cross-lingual homographs is relatively poor and lags far behind human performance. Meanwhile, we identify the phenomenon of **homograph shortcut**: the model tends to preferentially interpret the cross-lingual homographs in its easy-to-understand language, even when the meanings of these homographs differ entirely between the two languages. Most models are likely to commit the homograph shortcut on over 50% of homographs, with this figure reaching as high as 80.12% in the worst case. Since most homographs between Japanese and Chinese are synonyms, using homograph shortcuts is a reasonable approach. What we want to investigate here is whether it is possible to effectively handle cross-lingual homographs that are less frequent and semantically divergent, rather than synonymous.

The main contributions of this paper can be summarized as follows:

- We introduced a new dataset, Doppelganger-JC, which benchmarks the ability of LLMs to correctly handle homographs in cross-lingual tasks between Chinese and Japanese.
- Through the experiments using Doppelganger-JC, we pointed out that the issue of homograph shortcuts is prevalent across various models.
- We analyzed the potential causes of the homograph shortcut from the perspective of linguistic characteristics. We found that LLMs struggle to notice that the word is a cross-lingual homograph, especially when it shares the same part-of-speech (POS) between the two languages.

2 Related Work

Cross-lingual Homographs Dijkstra et al. (1999) defines cross-lingual homographs (they call "interlingual homographs" in the paper) as *"words in different languages share the same orthographic form"*. These words have mostly been investigated in the aspect of bilingual word recognition (Kennette and Van Havermaet, 2012; Dijkstra et al., 1999; Hoversten and Traxler, 2016).

[1] We use the word "doppelganger" because the words we deal with in this dataset have the same form on the surface, but they have completely different meanings, which resembles "doppelganger" — as in a ghost or shadow of yourself.

[2] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

[3] https://ai.meta.com/blog/meta-llama-3/

[4] https://telnyx.com/llm-library/mistral-7b-instruct-v0-2

Some papers discuss how language models handle cross-lingual homographs. Tanwar et al. (2025) conducts word pair disambiguation, semantic judgment, and semantic constraint sentences in English-Spanish and French-German and finds that LLMs perform worse than a random rate in cross-lingual homographs. Sterner and Teufel (2023) dealt with the task of determining where code-switching between German and English occurs and found that neural language models perform the best in disambiguating interlingual homographs. Nikov et al. (2024) does the task to identify whether the given cross-lingual word pair is a false friend pair or not. They find that the combination of BERT word embedding similarities and co-occurrence rate in the parallel corpus helps best to do false friend classification tasks.

However, previous works have not constructed datasets to focus on whether LLMs can correctly distinguish the meanings of cross-lingual homographs. Therefore, we construct a new benchmark dataset and conduct a detailed analysis using it.

Cross-lingual Homographs between Japanese and Chinese Japan was originally a society without writing, but then kanji characters were introduced about 1,600 years ago, and Japanese people started using them to write down Japanese (沖森卓也, 2010). In the process of Japanese characterization, some foreign words derived from Chinese underwent generalization, abstraction, and subdivision of meaning, changes in part-of-speech, and changes in nuance, and became established as words with meanings different from their original Chinese meanings (中川正之, 2013).

It is very important to construct the cross-lingual homograph dataset in Japanese and Chinese because the multilingualism of LLMs is more and more crucial, and the method to construct it can be applied to create cross-lingual homograph datasets for other language pairs.

3 Dataset Construction

To construct **Doppelganger-JC**, we followed a three-step pipeline which comprises: word-meanings set construction, example sentence generation, and task generation, as shown in Figure 2. The following subsections will elaborate on the three steps of constructing the benchmark.

3.1 Homograph Word-Meanings Set Construction

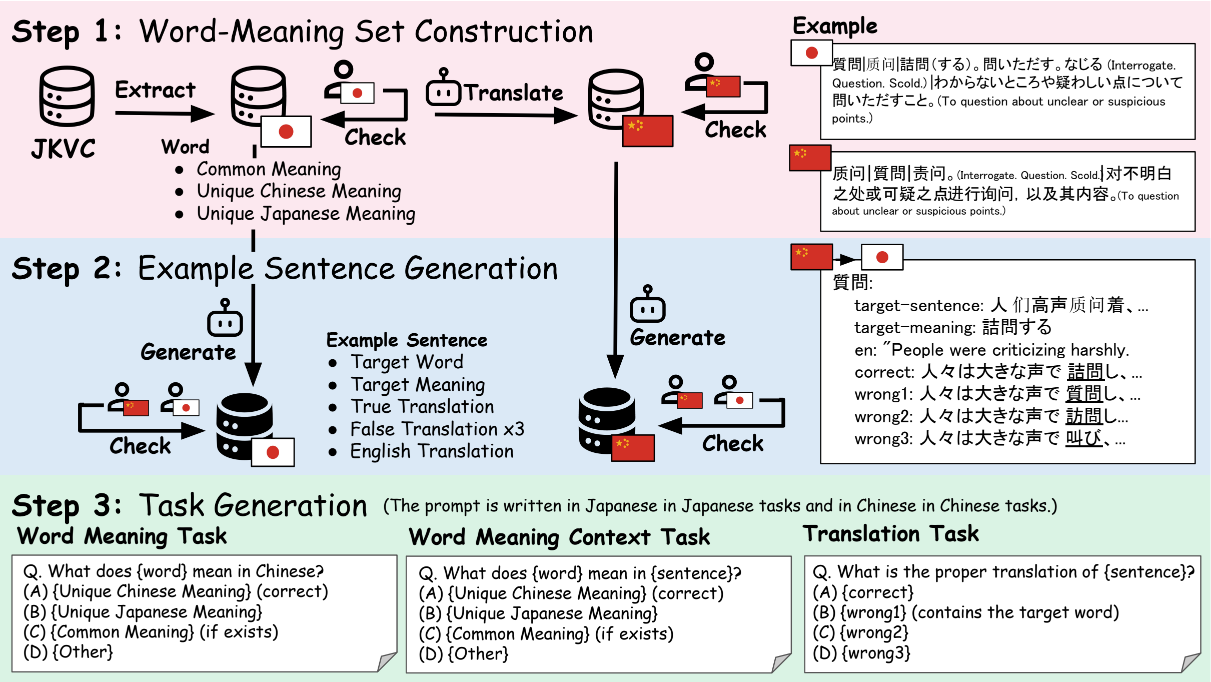
The initial step involves identifying the cross-lingual homographs that are shared between Chinese and Japanese. We refer to *the Database of Japanese Kanji Vocabulary in Contrast to Chinese* (JKVC) Version 2.00 (松下達彦 et al., 2020), which is a collection of Japanese-Chinese cross-lingual homographs. JKVC classifies words into the following types. **Type-1**: have completely different meanings; **Type-2**: share a common meaning, but also have unique meanings in respective languages; **Type-3**: have unique meanings only in Japanese; **Type-4**: have unique meanings only in Chinese; **Type-5**: not a word in Chinese; **Type-6**: have the same meanings. To facilitate understanding, we use Venn diagrams to demonstrate Types 1 to Type-6 in Figure 5 in Appendix A and give an example for each type in Table 9 in Appendix A.

Based on JKVC, we created a Chinese-Japanese homograph word-meanings set, which is a set of words and meanings prepared in both Japanese and Chinese, with the following steps: (1) We extracted all homographs of types 1 to 4 from JKVC with their meanings (see Table 9 in Appendix A) written in Japanese. We exclude types 5 and 6 homographs because they are less prone to misuse; (2) A native Japanese speaker performed checks on these homographs to confirm that all words had their meanings written in Japanese. In some cases, there was only a meaning written in Chinese provided for a word, and in such cases, we translate the meaning into Japanese. Also, if there are multiple meanings in the word, we use a character " ； " to separate the meaning; (3) The homograph word-meanings set was translated into Chinese using GPT-4.1[5], with prompts given in Table 10 in Appendix B; (4) A native Chinese speaker then performed checks on the translated homograph word-meanings set to confirm the accuracy of their translation. To ensure accuracy, annotators responsible for quality checking also used online dictionaries[6] as references.

The homograph word-meanings set ultimately records 1,290 cross-lingual homographs, along with their Chinese-Japanese common meanings, their Japanese-unique meanings, and their Chinese-unique meanings. It comprehensively covers a wide range of significant homographs in Japanese and Chinese, because our investigation of the

[5] https://openai.com/index/gpt-4-1/

[6] https://cjjc.weblio.jp/



- Common Meaning
- Unique Chinese Meaning
- Unique Japanese Meaning

Example Sentence

- Target Word
- Target Meaning
- True Translation
- False Translation x3
- English Translation

質問|质问|詰問(する)。問いただす。なじる (Interrogate. Question. Scold.)|わからないところや疑わしい点について問いただすこと。(To question about unclear or suspicious points.)

质问|質問|责问。(Interrogate. Question. Scold.)|对不明白之处或可疑之点进行询问，以及其内容。(To question about unclear or suspicious points.)

質問:
target-sentence: 人们高声质问着、...
target-meaning: 詰問する
en: "People were criticizing harshly.
correct: 人々は大きな声で 詰問し、...
wrong1: 人々は大きな声で 質問し、...
wrong2: 人々は大きな声で 訪問し...
wrong3: 人々は大きな声で 叫び、...

(The prompt is written in Japanese in Japanese tasks and in Chinese in Chinese tasks.)

Q. What does {word} mean in Chinese?
(A) {Unique Chinese Meaning} (correct)
(B) {Unique Japanese Meaning}
(C) {Common Meaning} (if exists)
(D) {Other}

Q. What does {word} mean in {sentence}?
(A) {Unique Chinese Meaning} (correct)
(B) {Unique Japanese Meaning}
(C) {Common Meaning} (if exists)
(D) {Other}

Q. What is the proper translation of {sentence}?
(A) {correct}
(B) {wrong1} (contains the target word)
(C) {wrong2}
(D) {wrong3}

Figure 2: The pipeline to construct Doppelganger-JC.

Japanese-English-Chinese basic sentence data[7], which contains 5,304 typical translation pairs, reveals that 603 sentences included Type 1 homographs from our dataset, with similar numbers present for Types 2 to 4.

The homograph word-meanings set is available in both Japanese (JA) and Chinese (ZH) versions, and the distribution of homographs across various types is presented in Table 2. Notably, the Japanese version does not include Type-4 homographs, as they do not possess unique meanings in Japanese, so misuse of this type does not occur in the Japanese context. The Chinese version of the homograph word-meanings set does not include Type-3 for the same reason.

3.2 Example Sentences Generation

To better analyze whether LLMs can correctly use and understand these words in contexts with different languages, we generated one example sentence for each word we collected in the word-meanings set. Throughout this subsection, we illustrate with the generation of Japanese example sentences. The procedure for generating the Chinese version is exactly the same as for Japanese.

We prompted GPT-4.1 to generate sentences using words from a homograph word-meanings set, ensuring that the meaning of the target word in each sentence corresponds to one of its unique meanings in Japanese. We also generated the correct Chinese translation, three distorted Chinese translations (one of them must include the target word itself in the translation), and an English translation of the Japanese sentence to facilitate task generation. The example sentences in Chinese and corresponding translations were also generated through the same process. The prompt used for generating the example sentences is given in Table 11 in Appendix C.

We would like to elaborate further on the distorted translation generations. Since we require that the target word in the Japanese example sentences must possess a meaning unique to Japanese, i.e., its meaning in Japanese must differ from its meaning in Chinese, we can therefore conclude that any Chinese translation containing the target word is necessarily incorrect, as exemplified in Figure 1. If the model selects a homograph-included distraction in a translation task, or if the open-ended translated sentence includes the target homograph, we refer to this phenomenon as the "**homograph shortcut**".

To guarantee quality, native Japanese and Chinese speakers were asked to review the example sentences and their corresponding translations. The evaluation criteria included: (1) The example sentences and their correct Chinese translations must be fluent and natural; (2) The target word in each

[7] https://nlp.ist.i.kyoto-u.ac.jp/?日英中基本文データ#p8f18e6d

example sentence must correspond to one of its unique meanings in Japanese; (3) The correct Chinese translation must not include the target word; (4) Among the distraction Chinese translations, at least one of them should contain the target word. Criteria (3) and (4) are adopted to test homograph shortcuts.

3.3 Task Design

We aim to evaluate whether LLMs can accurately understand and utilize cross-lingual homographs through the Doppelganger-JC benchmark. Therefore, we have designed three types of tasks: word meaning task, word meaning in context task, and sentence translation task. In order to facilitate the observation of error types, we carefully designed the distraction options and adopted the Multiple-Choice Question Answering (MCQA) format as the primary form for these tasks. In this subsection, we still take Japanese as an example. The task for Chinese is identical to that for Japanese, with the only difference being the language itself.

Word Meaning Task In the meaning task, LLMs are required to select the meaning of a given word from the homograph word-meanings set in Japanese. The three distraction options consist of the unique meanings of this word in Chinese (which should differ from its unique meanings in Japanese), as well as two completely different meanings, which are in the word-meaning list that are one or two indices away from the target word. The prompt is shown in Table 12.

Word Meaning in Context Task The word meaning in context task requires the LLMs to select the meaning of a given word as it appears in its example sentence. The design of the distraction options is identical to that in the word meaning task. The prompt is shown in Table 13.

Translation Task The translation task requires the LLMs to select the correct Chinese translation for each example sentence. The design of the distracting Chinese translations has already been detailed in § 3.2. As stated before, at least one of the distraction options contains the target homograph. The distraction options that include the target word are likely to exhibit higher similarity to the correct translation, as shown in Figure 3. In subsequent sections, we will further analyze which type of distraction option LLMs tend to select. The prompt is shown in Table 14.

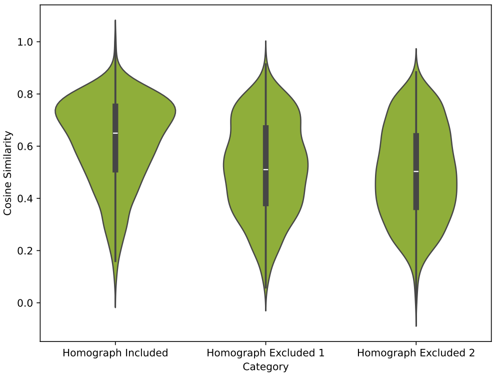


Figure 3: Cosine similarity between the distraction options and the correct translation in Japanese-to-Chinese translation tasks.

Lang	Type-1	Type-2	Type-3	Type-4	Total
JA	464	182	355	-	1,001
ZH	464	182	-	289	935

Table 2: The number of Type-1 to Type-4 homographs in the homograph word-meanings set in Japanese (JA) and Chinese (ZH).

4 Experiments

4.1 Experiments Settings

Models We selected representative open-source LLMs from each of the three categories to evaluate their performance on Doppelganger-JC, as shown in Table 1. We assume that LLMs will demonstrate superior performance in the language of their respective countries. We used one Tesla V100 GPU.

All models, tokenizers, and libraries used in this study are open-source and licensed under permissive terms. This ensures that our experimental setup is fully reproducible and freely reusable for academic purposes. We conducted experiments with various parameter sizes as well, and report the results in Appendix N.

Human Baseline To facilitate comparison, we recruited two Chinese international student volunteers currently studying in Japan to participate in benchmark testing. Both participants are native speakers of Chinese and have demonstrated Japanese proficiency at the JLPT-N2 [8] level. From each type of homograph in the tasks, we sampled 50% of the questions and allocated them to the two volunteers. The volunteers were given the same instructions as LLMs to answer the questions as

[8] https://www.jlpt.jp/e/about/index.html

shown in Fig. 2. We calculated their average accuracy on these questions as a representation of human performance.

Evaluation Metrics For the three tasks, we prompted the LLM with different instructions and obtained the LLM's predicted answers by selecting the option with the lowest perplexity. We use accuracy to evaluate the LLMs' performance on our benchmark.

4.2 Main Results

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-3	T-1	T-2	T-4
Human	95.04	93.40	93.40	94.47	92.20	95.28	95.82	93.82
llm-jp	**64.87**	**50.55**	**79.00**	49.17	**58.36**	65.49	**72.63**	73.08
ELYZA-JP	53.88	42.86	74.24	**51.38**	55.81	60.44	67.60	70.63
Qwen2.5-7B	30.60	25.27	59.09	44.20	27.20	64.18	70.39	**73.78**
Baichuan2-7B	20.91	18.68	57.14	44.20	46.74	58.90	62.57	64.34
Llama-3.1	41.16	34.62	67.75	45.30	46.74	**70.55**	71.51	72.38
gemma-7b	41.38	35.71	64.94	46.41	46.46	60.00	67.04	68.53
Mistral-7B	24.35	17.58	51.30	37.02	39.09	41.32	46.37	45.80

Table 3: Performance (%) of LLMs on the Japanese word meaning task, word meaning in context task, and Chinese-to-Japanese translation task.

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-4	T-1	T-2	T-3
Human	94.40	90.66	96.10	92.73	98.84	94.59	95.03	90.94
llm-jp	28.45	37.91	40.66	30.17	36.36	58.23	61.88	50.99
ELYZA-JP	24.14	29.67	47.47	35.75	50.70	52.81	58.56	48.16
Qwen2.5-7B	**62.93**	**54.40**	**72.97**	**48.04**	**53.85**	69.70	72.38	61.47
Baichuan2-7B	60.34	51.10	71.21	54.19	66.78	61.69	67.96	50.14
Llama-3.1	49.57	47.80	63.96	46.93	45.45	**72.08**	**76.24**	**66.57**
gemma-7b	53.66	51.65	56.70	39.66	47.20	61.04	65.75	54.39
Mistral-7B	44.83	42.86	53.85	42.46	47.90	49.78	56.35	42.78

Table 4: Performance (%) of LLMs on the Chinese word meaning task, word meaning in context task, and Japanese-to-Chinese translation task.

Tables 3 and 4 present the performance of various LLMs on Doppelganger-JC.

Although human evaluators (native Chinese speakers with a certain level of Japanese proficiency) are able to perform all tasks with a high accuracy rate, the accuracy rate of LLM remains low. Especially, from the perspective of the task, the meaning task is the most challenging. Without sufficient context and relying solely on simple prompts, it is difficult for models to accurately identify the meanings of cross-lingual homographs in the specified language. In some cases, model performance is even lower than that of a random baseline. Once the context is provided, i.e., word meaning in context and translation tasks, the model's performance is improved. However, there remains a noticeable gap compared to the human baseline. This indicates that accurately understanding and handling cross-lingual homographs continues to pose a significant challenge for LLMs.

From the perspective of task language, for monolingual tasks, i.e., meaning tasks and meaning-in-context tasks, models developed for Japanese and Chinese demonstrate superior performance on tasks in their respective languages. Another finding is that the performance of Chinese models on Japanese tasks or Japanese models on Chinese tasks is even worse than that of models from other countries. We attribute this to the presence of a native language bias in both Chinese and Japanese models. For cross-lingual homographs, these models may tend to wrongly select the meaning associated with the word in their respective native languages. Conversely, when dealing with cross-linguistic translation tasks, the presence of native language bias in Chinese and Japanese models negatively impacts their performance, resulting in Llama 3.1, which is from neither China nor Japan, outperforming them.

From the perspective of word type, in monolingual tasks, LLMs generally exhibit the lowest accuracy in understanding the meanings of Type-2 words. We hypothesize that this phenomenon arises because Type-2 words tend to have a wider range of possible meanings, as shown in Table 9 in Appendix A. Thus, it is more challenging for the model to accurately capture the unique meanings of these words in either Chinese or Japanese. As a comparison, Table 17 in Appendix I reports results for the tasks with Type-6 words, which share the same meaning in both Chinese and Japanese.

In summary, the performance of LLMs varies across different tasks, task languages, and word types; however, there remains a substantial gap compared to the human baseline. This suggests that the models exhibit a considerable number of misuses in Chinese-Japanese cross-lingual homographs. In the following sections, we will further analyze the characteristics of such misuses.

5 Analysis

5.1 Analysis of Meaning and Meaning in Context Tasks

Context Robustness of Word Meanings As shown in § 4.2, LLMs are less likely to commit homograph shortcuts on the meaning tasks when the context is provided. Therefore, we hope to investigate the context robustness of word meanings in different models.

We calculated the cosine similarities between the final-layer embeddings of the target word with and without the context. A higher degree of similarity indicates that the model's understanding of the word is more robust and less dependent on context. We used the median as a threshold to divide the words into high-robustness and low-robustness groups. Then we calculated the accuracy of meaning tasks with and without context for each group. The result is demonstrated in Table 5.

Model	Chinese				Japanese			
	w/ context		w/o context		w/ context		w/o context	
	high	low	high	low	high	low	high	low
llm-jp	0.39	0.36	0.33	0.29	0.78	0.63	0.68	0.54
ELYZA-JP	0.45	0.43	0.26	0.26	0.74	0.62	0.58	0.43
Qwen2.5-7B	0.68	0.64	0.62	0.60	0.53	0.56	0.33	0.25
Baichuan2-7B	0.66	0.67	0.59	0.56	0.57	0.50	0.24	0.16
Llama-3.1	0.59	0.59	0.50	0.48	0.62	0.61	0.43	0.36
gemma-7b	0.54	0.50	0.56	0.50	0.59	0.61	0.41	0.39
Mistral-7B	0.50	0.51	0.43	0.45	0.50	0.45	0.24	0.22

Table 5: Performance of LLMs on the word meaning tasks, with words divided into high robustness and low robustness groups.

The high-robustness group is characterized by minimal sensitivity to contextual variation, indicating that these models remain faithful to the lexical semantics acquired during pre-training. Consequently, they exhibit strong performance on the without-context task, particularly for languages that were prioritized in their pre-training corpora. This observation supports an argument that Japanese models demonstrate higher scores for Japanese tasks when context is not given, whereas Chinese models show the opposite pattern.

In contrast, the low-robustness group appears more capable of modulating word meaning in response to contextual cues, which is advantageous for the with-context task. An analysis of the score difference between the with-context and without-context settings suggests that the low-robustness group generally achieves larger gains, providing evidence for its stronger context adaptability.

5.2 Analysis of Translation Tasks

Proportion of Homograph Shortcut As discussed in § 3.3, the translation task involves two categories of distraction options: ones including the target homographs and ones excluding them. If the model selects a homograph-included distraction, a homograph shortcut occurs. We seek to explore the proportion of homograph shortcuts.

Table 6 presents the proportion of homograph shortcuts among all errors. It can be observed that the primary errors made by the model are due to homograph shortcuts, i.e., the model shows a tendency to prefer the homograph-included distraction option. This indicates that the model does not have an adequate understanding of cross-lingual homographs and tends to mistakenly treat homographs in different languages as the same word.

Model	Japanese-to-Chinese	Chinese-to-Japanese
llm-jp	72.87	67.84
ELYZA-JP	84.42	74.53
Qwen2.5-7B	84.36	81.44
Baichuan2-7B	85.89	71.91
Llama-3.1	67.24	65.53
gemma-7b	81.14	78.25
Mistral-7B	68.23	75.10

Table 6: The proportion of errors (%) in which the LLM selects a homograph-included distraction option.

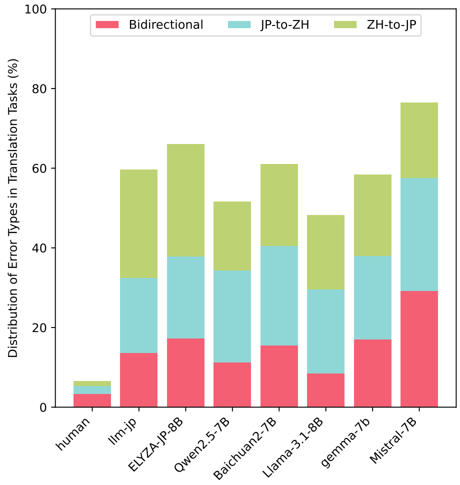


Figure 4: Distribution of homograph misuse types in translation tasks. Color blue indicates that the model misuses the homograph only from Japanese to Chinese (JA-to-ZH) translation, while green indicates only from Chinese to Japanese (ZH-to-JA), and red indicates bidirectional misuse.

Directionality of Homograph Misuse An important question is whether the misuse of cross-lingual homographs in translation tasks demonstrates directionality, i.e., for a given word, does the model misuse it only when translating from Chinese to Japanese, only from Japanese to Chinese, or in both directions?

Figure 4 demonstrates the distribution of homograph misuse types in translation tasks. It can be observed that humans also exhibit instances of homograph misuse; however, 52.5% of these cases are bidirectional. In other words, when misusing homographs, humans tend to assume that the word carries the same meaning in both languages (as we see in Figure 1). However, the performance of LLMs is notably different. The extent of bidirectional misuse in LLMs is relatively limited. Furthermore, Japanese models tend to make more mistakes in Chinese-to-Japanese translation tasks, whereas models for other languages display the opposite trend. We hypothesize that this phenomenon arises because the training corpus for the Japanese model contains a higher proportion of Japanese texts, leading the model to more readily interpret cross-lingual homographs in Chinese contexts according to their Japanese meanings, and vice versa.

Open-Ended Translation The preceding analyses were all conducted using multiple-choice tasks. We are also interested in examining, in open-ended translation, how frequently the LLM erroneously commits homograph shortcuts. Therefore, we further carried out open-ended translation experiments.

For the open-ended translation task, we introduce a new criterion, misused homographs at k samples ($MH@k$). For each homograph in the word-meanings set, we sample k candidate translations for its corresponding example sentence. $MH@k$ is defined as the proportion of homographs in the word-meanings set for which the LLM commits a homograph shortcut at least once out of the k translation samples. We count as homograph shortcuts regardless of whether they are used in Japanese or Chinese spelling. An increase in a model's $MH@k$ score indicates a higher likelihood of homograph shortcut across a wider spectrum of homographs.

Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=1$	$k=3$	$k=5$	$k=10$
	Japanese-to-Chinese				*Chinese-to-Japanese*			
llm-jp	31.67	43.56	48.25	54.15	**22.46**	**29.73**	**33.26**	**37.43**
ELYZA-JP	54.15	73.63	80.12	86.61	32.51	51.98	58.40	66.42
Qwen2.5-7B	**22.88**	**27.87**	**30.07**	**32.07**	26.84	38.50	42.46	45.88
Llama-3.1	31.17	45.75	50.05	58.64	29.30	40.64	45.56	50.70
Mistral-7B	31.17	41.76	45.35	50.45	28.45	42.03	47.91	54.65

Table 7: $MH@k \downarrow$ ($k = 1, 3, 5, 10$) scores (%) of representative models in open-ended translation tasks. The best-performing model is indicated in bold, while the second-best is underlined.

Table 7 presents the $MH@k$ scores for several representative models. It can be observed that Qwen-2.5 demonstrates relatively stronger performance on the Japanese-to-Chinese task, while both llm-jp and Qwen-2.5 perform comparatively well on the Chinese-to-Japanese task. For the other models, $MH@10$ exceeds 50%, indicating a high likelihood of homograph shortcut on more than half of the homographs.

We also compiled a list of homographs that were most frequently misused by all models, which are presented in Table 15 in Appendix G. It can be observed that most of these words share the same part of speech (POS) in both Chinese and Japanese. We hypothesize that, in such cases, LLMs may find it more challenging to distinguish the different meanings of these words in Chinese and Japanese, since their grammatical roles within the sentence are identical. We will elaborate on this point in the following paragraph.

Model	**POS-same**		**POS-different**	
	Z→J	**J→Z**	**Z→J**	**J→Z**
llm-jp	55.98	55.63	74.72	59.60
ELYZA-JP	52.45	50.07	72.24	55.77
Qwen2.5-7B	58.31	61.51	73.76	74.27
Baichuan2-7B	46.91	56.44	67.89	66.23
Llama-3.1	61.52	70.11	71.53	72.86
gemma-7b	54.00	54.31	73.54	64.99
Mistral-7B	28.92	45.80	49.53	54.42

Table 8: Performance (%) of LLMs on the translation task for POS-different words and POS-same words from Type-1 and Type-2. In this table, **J** represents Japanese while **Z** represents Chinese.

POS of Cross-lingual Homographs By looking at the words that are most frequently misused, we found that cross-lingual homographs sharing the same POS are more prone to misuse. Therefore, we categorize these homographs based on whether their POS is the same in both languages, and examine the performance of LLMs on each category. We extract POS of homographs using a Japanese POS and morphological analyzer, MeCab[9], and a Chinese one, jieba[10]. Table 8 presents the model's performance on translation tasks involving two categories of words. The results demonstrate that LLMs are more adept at distinguishing the different functions and usages of POS-different words across the two languages. Although LLMs do not achieve human-level performance on POS-different words, we believe leveraging linguistic features such as POS may represent one potential approach to addressing cross-lingual homograph misuse in models.

[9] https://taku910.github.io/mecab/

[10] https://github.com/fxsjy/jieba

6 Conclusion

In this study, we introduced a new benchmark, Doppelganger-JC, to systematically evaluate how well LLMs handle cross-lingual homographs between Chinese and Japanese. Our benchmark comprises three types of tasks: word meaning, word meaning in context, and sentence translation.

Our results revealed a significant phenomenon, homograph shortcut, where LLMs tend to interpret cross-lingual homographs in their easy-to-understand language even when their meanings differ completely across languages.

Further analysis suggested that the likelihood of a homograph shortcut increases when the homograph shares the same part-of-speech (POS) in both languages. When the POS differs, models are better able to recognize the semantic difference, likely due to syntactic constraints that aid disambiguation.

Overall, our findings demonstrate that handling cross-lingual homographs remains a major challenge for LLMs. Addressing this issue will require not only curated data but also new modeling approaches that can better leverage linguistic features like POS to detect and resolve the ambiguity.

Limitations

The main limitations of this paper are as follows, and we aim to address them in our future work:

The homograph word-meanings set presented in this paper does not encompass all cross-lingual homographs between Chinese and Japanese, due to constraints imposed by the source word-meanings set JKVC. Moreover, the usage of vocabulary in JKVC tends to be relatively traditional and formal. The development of the Internet has had a substantial impact on the change of word meanings, and the misuse of cross-lingual homographs may now present in ways different from the past.

Despite its high quality resulting from human annotation, the dataset size is not sufficient to support the fine-tuning of large-scale language models. As a next step, we intend to increase the level of automation in our data generation process, enabling faster creation of datasets containing cross-lingual homographs. Subsequently, we aim to introduce a cross-lingual-homograph-aware loss function, allowing the model to better distinguish the semantic variations of identical words across different languages and contexts.

Also, we adopted only MCQA in this paper, but it can also be extended to translation generation and disambiguation tasks using alternative methods (e.g., LLM-as-a-judge). Also, we can test mitigation methods such as fine-tuning, prompting, or architectural changes.

We analyzed cross-lingual homographs solely between Chinese and Japanese. However, the phenomenon of LLMs misusing cross-lingual homographs occurs in many other languages, such as among Latin-alphabet languages and Slavic languages. We hope that our approach can be applied to other languages to identify and address similar issues.

Ultimately, the results of human evaluation are believed to depend on language ability. Currently, the number of human evaluators is limited, and there is a possibility of bias, so further larger-scaled analysis from the perspective of linguistic is desired.

Acknowledgments

We would like to express our gratitude to Tatsuhiko Matsushita, one of the developers of the JKVC (Database of Japanese Kanji Vocabulary in Contrast to Chinese), for kindly allowing us to use their data, and to him and Hitomi Kinoshita for their valuable discussions.

References

Ton Dijkstra, Jonathan Grainger, and Walter J.B. van Heuven. 1999. Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, 41(4):496–518.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. elyza/llama-3-elyza-jp-8b.

Liv J Hoversten and Matthew J Traxler. 2016. A time course analysis of interlingual homograph processing: Evidence from eye movements. *Bilingualism: Language and Cognition*, 19(2):347–360.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, and 1 others. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. *arXiv preprint arXiv:2410.05873*.

Lynne N Kennette and Lisa R Van Havermaet. 2012. Interlingual homograph recognition by bilinguals: A new paradigm. *The New School Psychology Bulletin*, 9(2):7–16.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, and 64 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *Preprint*, arXiv:2407.03963.

Mitko Nikov, Žan Tomaž Šprajc, and Žan Bedrač. 2024. Cross-lingual false friend classification via llm-based vector embedding analysis. *Proceedings of the 10th Student Computing Research Symposium (SCORES' 24)*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Igor Sterner and Simone Teufel. 2023. Tongueswitcher: Fine-grained identification of german–english code-switching. Association for Computational Linguistics.

Eshaan Tanwar, Gayatri Oke, and Tanmoy Chakraborty. 2025. Multilingual llms struggle to link orthography and semantics in bilingual word processing. *arXiv preprint arXiv:2501.09127*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Qwen Team. 2025. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025a. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Zhen Yang, Ping Jian, and Chengzhi Li. 2025b. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917, Albuquerque, New Mexico. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

中川正之. 2013. 漢語からみえる世界と世間: 日本語と中国語はどこでずれるか. 岩波書店.

松下達彦, 陳夢夏, 王雪竹, and 陳林柯. 2020. 日中対照漢字語データベースの開発と応用. 日本語教育, 177:62–76.

沖森卓也. 2010. はじめて読む日本語の歴史: うつりゆく音韻・文字・語彙・文法. ベレ出版.

A Types of Cross-Lingual Homographs

Figure 5 demonstrates the definition of different types of homographs according to JKVC, and Table 9 gives an example for each type.

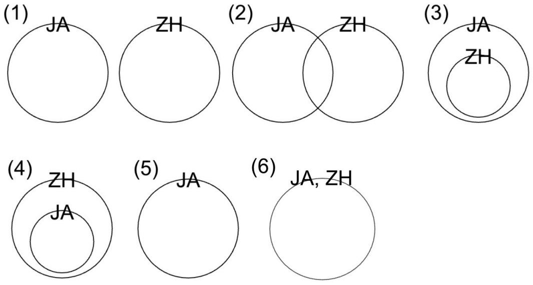


Figure 5: Venn diagrams of different types of homographs according to JKVC.

B Prompts Used for Translating the Homograph Word-Meanings Set

Table 10 shows the prompt used to obtain the draft Chinese version of the homograph word-meanings set.

C Prompts for Example Sentences Generation

Table 11 shows the prompt used to obtain the example sentences and distraction options.

Type	Definition	Example	Meaning		
			Common	JA Unique	ZH Unique
1	Have completely different meanings.	勉強	-	To study.	Reluctant.
2	Share a common meaning, but also have unique meanings in respective languages.	主人	The host.	The husband.	The owner.
3	Have unique meanings only in Japanese.	時間	Time.	An hour.	-
4	Have unique meanings only in Chinese.	幾何	Geometry.	-	How many.
5	Not a word in Chinese.	病院	-	A hospital.	-
6	Have same meanings.	社会	The society.	-	-

Table 9: Definition of the six cross-lingual homograph types and examples of each type.

Prompt
Please translate the following in Chinese and output it in a single line:

Table 10: Prompts used for translating the homograph word-meanings set into Chinese.

D Prompts to Measure Perplexities

To get perplexities for the three tasks in this paper, we use the following prompts: word meaning tasks (Table 12), word meaning in context tasks (Table 13), and translation tasks (Table 14).

E Cosine Similarities of Options

In the main text, similarity between the distraction options and the correct translation in Japanese-to-Chinese translation tasks is presented in Figure 3; the result for Chinese-to-Japanese translation is presented in Figure 6. Also, similarity between the correct meaning and other options in the meaning task is presented in Figure 7 and 8.

F Performance of LLMs on Doppelganger-JC

Figure 9 shows Table 3 and 4 in graph form. In the figure, blue bars represent Japanese LLMs (llm-jp and ELYZA-JP), while pink bars represent Chinese LLMs (Qwen-2.5-7B and Baichuan-7B). Green and orange bars correspond to models trained in other languages.

Prompt
The word "{word}({word_zh})" has the following meanings in Japanese and Chinese. Japanese: {jp_mean} Chinese: {zh_mean} By using the meaning in Japanese, create a Japanese sentence that could be misinterpreted in Chinese, and provide three options for its Chinese translation. Note that the output should be in JSON format, as follows: "{word_zh}": {{ "target-sentence": "Japanese sentence". "target-meaning": "the meaning of the word here", "en": "English translations of the sentence", "correct": "Correct Chinese translation", "wrong1": "Wrong Chinese translation(Note that this sentence must contain the word "{word_zh}")", "wrong2": "Wrong Chinese translation", "wrong3": "Wrong Chinese translation" }} ## Examples

Table 11: Prompts used for generating the example sentences.

Lang	**Prompt**
ZH	"{word}"这个词的意思是"{meaning}".
JA	「{word}」という単語の意味は「{meaning}」です。

Table 12: Prompts used for word meaning tasks.

Lang	Prompt
ZH	"{sentence}"这句话中使用的"{word}"一词的含义是"{meaning}".
JA	「{sentence}」という文章内で使われている「{word}」の意味は「{meaning}」です。

Table 13: Prompts used for word meaning in context tasks.

Lang	Prompt
ZH	"{original}" 的中文翻译是"{translated}".
JA	"{original}"という中国語の文章の日本語訳は「{translated}」です。

Table 14: Prompts used for translation tasks.

G Frequent Misused Words

Table 15 shows a list of cross-lingual homographs that were most frequently misused by all models. Also, Figure 10 shows the examples of actual model errors.

H MCQA Bias toward Option A

Other than the perplexity-based approach, we prompted the LLM with different instructions and obtained the LLM's predicted answers through directly responding to the option for the three tasks in this paper (i.e., meaning tasks, meaning in context tasks, and translation tasks). The direct response method can be applied to any model, including closed-source models such as GPT-4. However, our experiments in Table 16 revealed a bias toward choosing option A when using this method, a phenomenon also noted in some previous studies (Zheng et al., 2023; Yang et al., 2025b; Pezeshkpour and Hruschka, 2023). We reorder the options so that the answer becomes A, B, C, D (which means that we do the same task four times), and count how many times LLMs answer each option. Table 16 shows the percentage of each option that was selected. Almost all models, even GPT-4.1, answer "A" the most, and surprisingly, more than 60% in llm-jp and Mistral.

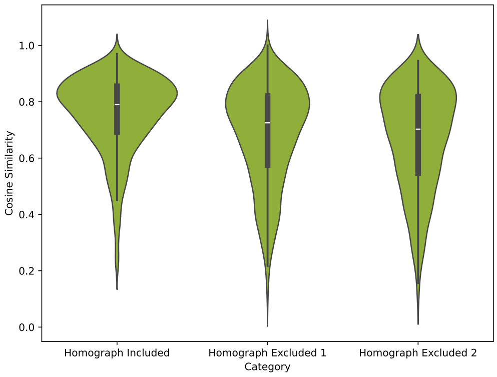


Figure 6: Cosine similarity between the distraction options and the correct translation in Chinese-to-Japanese translation tasks.

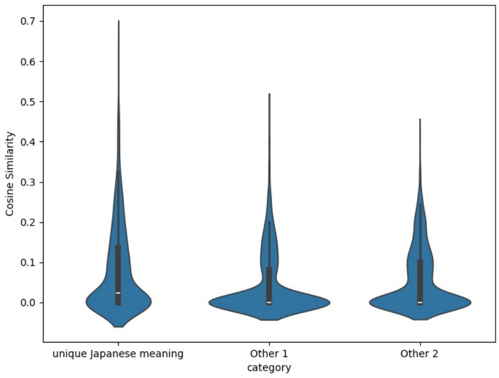


Figure 7: Cosine similarity between the correct meaning and other options in Chinese meaning tasks.

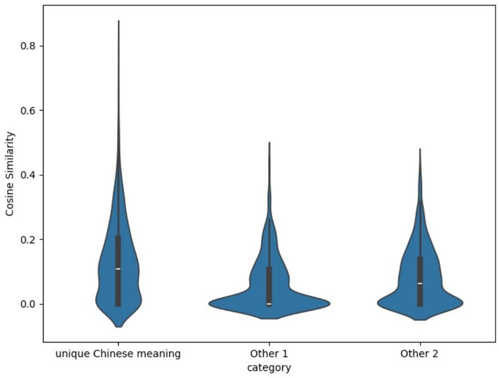


Figure 8: Cosine similarity between the correct meaning and other options in Japanese meaning tasks.

I The Baseline Result with Homographs

In this research, we focus only on the cross-lingual homographs and conclude that LLMs are not good at dealing with homographs. Is it truly correct? We use homophones (i.e., words which have the same form and the same meaning in two languages) as

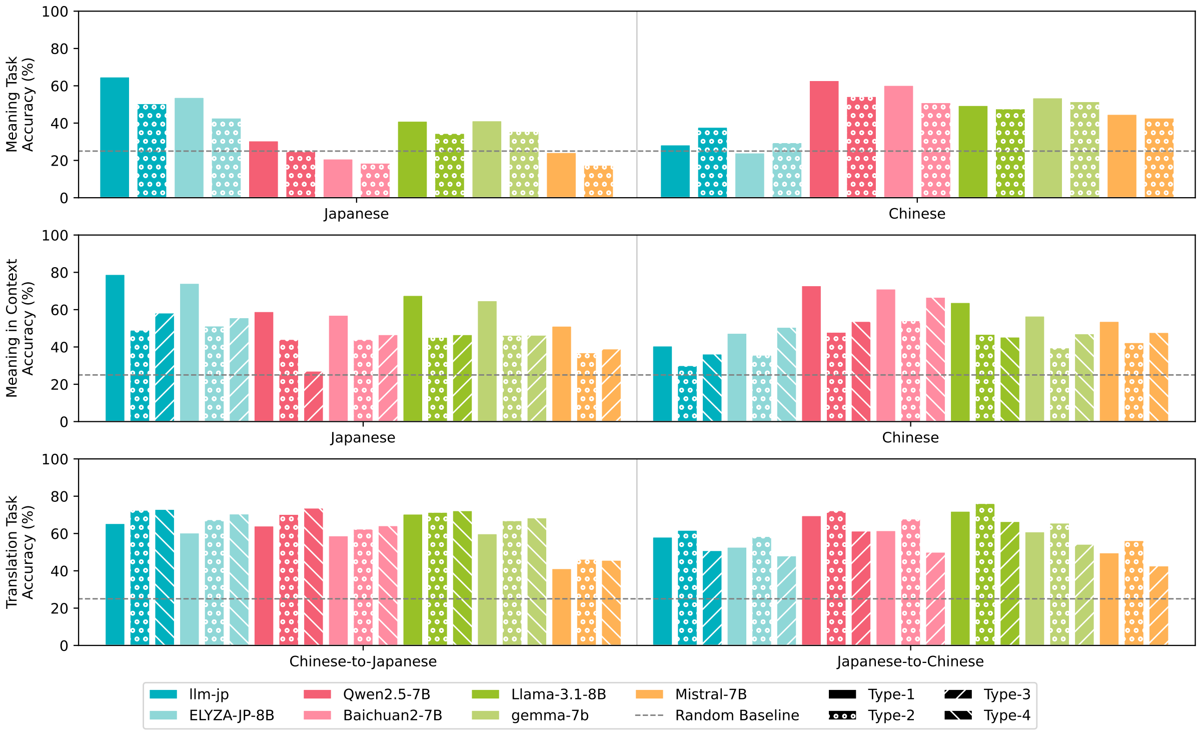
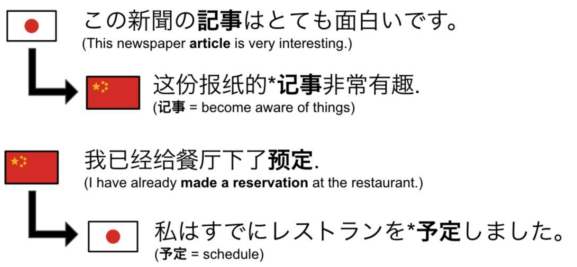


Figure 9: Performance of LLMs on Doppelganger-JC.



この新聞の**記事**はとても面白いです。
(This newspaper **article** is very interesting.)

这份报纸的***记事**非常有趣.
(**记事** = become aware of things)

我已经给餐厅下了**预定**.
(I have already **made a reservation** at the restaurant.)

私はすでにレストランを***予定**しました。
(**予定** = schedule)

Figure 10: Examples of actual model errors in translation tasks.

the baseline to ensure that LLMs perform worse in cross-lingual homographs.

JKVC provides 5,784 homophones as Type-6. We collect the top 100 nouns of them and adopt the pipeline method shown in § 3 to construct the baseline dataset. Table 17 shows the performance of each model on this baseline dataset.

Comparing the result with Table 3 and 4, the accuracies with the baseline dataset are better in all models and tasks. Therefore, we can still conclude that LLMs are not good at dealing with cross-lingual homographs.

In homophones, it is correct to commit the homograph shortcut. This fact makes it difficult for models to judge whether they can directly use the word in their translation or not. Also, Japanese models and Chinese models perform better than other English-centric models. It supports our hypothesis that it makes it easier for models to deal with words when they exist in both Chinese and Japanese, and the spelling and the meaning are the same.

J Distribution of Homograph Misuse Types

Table 18 shows the confusion matrices of translation accuracy. When we look at the case where one is correct and the other is wrong, we can see that Japanese models perform better at Chinese-to-Japanese tasks, and Chinese and other models perform better at Japanese-to-Chinese tasks. This supports the hypothesis that Japanese models are better at Japanese tasks and vice versa (§ 4.2) because the prompt is written in Japanese in the Chinese-to-Japanese translation task and Chinese in the Japanese-to-Chinese translation task.

K The Difference between With and Without Context

In § 5.1 we discuss the effect of context by using the models' final layers. Figure 11 and 12 show the results of all layers.

As a whole, words with high cosine similarity

Rank	**Word**	**Type**	**Meaning**		
			Common	**JA Unique**	**ZH Unique**
			Japanese-to-Chinese		
1	建立	1	-	To develop relationships.	To establish (an organization).
2	的	2	Transform some nouns into adjectives.	A target device used for archery or firearms shooting practice.	Indicate ownership.
3	忘年	3	Forget about the age gap.	Forget the hardships of the past year.	-
4	府	2	The government.	Osaka.	The house.
5	調整	3	Adjust to the correct state.	Dispatch and organize supplies.	-
			Chinese-to-Japanese		
1	本	1	-	A counter word used for slender items.	A counter word used for books or notes.
2	文章	4	An article.	-	The hidden meaning of words.
3	追	4	Chase.	-	Press for responsibility; to court (someone).
4	運	4	Luck.	-	Carry something from one place to the other.
5	中学	4	Junior high school.	-	Senior high school.

Table 15: Definition of the six cross-lingual homograph types and examples of each type.

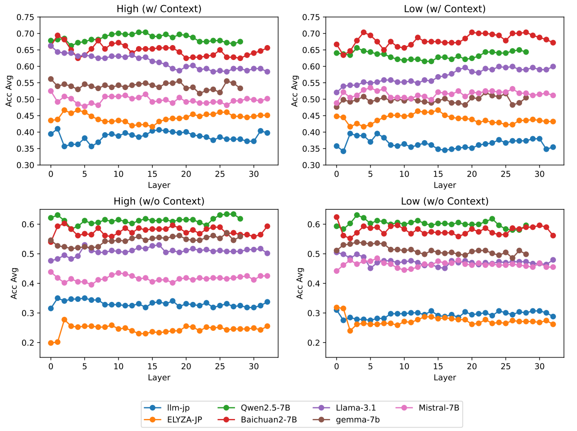


Figure 11: In each layer, words were split into high and low similarity groups based on the median cosine similarity. The table presents the average correctness (0 or 1) for each group in Chinese tasks, both with and without contextual information.

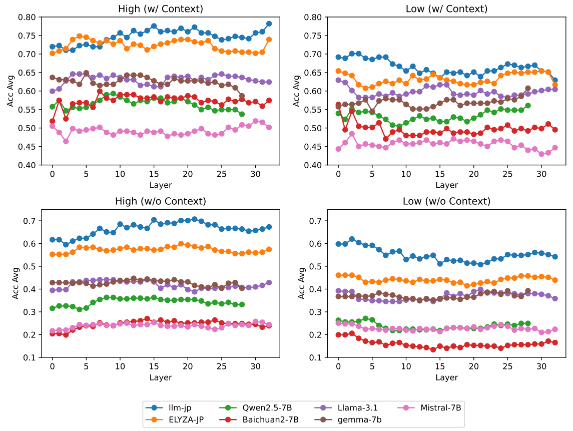


Figure 12: In each layer, words were split into high and low similarity groups based on the median cosine similarity. The table presents the average correctness (0 or 1) for each group in Japanese tasks, both with and without contextual information.

in the embeddings are more likely to be correct for the tasks in any layer because the average accuracy in each layer of the cosine-similarity-high group is higher than that of the cosine-similarity-low group when we compare graphs lined up side by side. Therefore, it can be assumed that it is possible to handle such words as cross-lingual homographs effectively from the early stages.

L More Detailed Data Related to MH@k

In § 5.2, we discuss $MH@k$, and Table 7 shows the proportion of homographs in the word-meanings

Model	Japanese								Chinese							
	Word Meaning				Translation				Word Meaning				Translation			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
llm-jp	23.96	**26.01**	23.68	26.35	**58.08**	20.31	9.11	12.50	**42.69**	25.56	11.37	20.38	**68.01**	11.52	9.12	11.35
ELYZA-JP	**27.41**	26.25	23.65	22.69	**42.65**	27.94	16.49	12.93	**45.87**	28.40	13.75	11.99	**52.63**	23.09	14.59	9.69
Qwen2.5-7B	**35.50**	23.37	20.62	20.50	**40.34**	22.99	19.83	16.84	**41.75**	24.44	17.31	16.50	**44.21**	21.88	17.36	16.55
Llama-3.1	**28.64**	26.35	22.68	22.33	**28.77**	28.69	21.81	20.73	**42.61**	26.55	14.59	16.25	**44.23**	23.20	16.07	16.51
Mistral-7B	**52.79**	12.50	17.03	17.69	**26.03**	29.74	23.07	21.16	**60.54**	12.84	12.45	14.17	**38.01**	24.01	19.36	18.62
GPT-4.1	**31.07**	24.22	22.12	22.59	**32.91**	24.05	21.69	21.36	**41.29**	25.81	15.98	16.91	**33.67**	24.63	21.26	20.44

Table 16: The LLM's bias for MCQA options. This shows the percentage of each option that appeared when the same question was given to the LLM four times, with the options rearranged so that A, B, C, and D were all correct answers. Ideally, all options should appear at a rate of 25%. Among each task, the option with the highest output ratio is indicated in bold.

Model	Japanese			Chinese		
	WM	WMC	T	WM	WMC	T
llm-jp	77.0	71.0	99.0	49.0	58.0	95.0
ELYZA-JP	59.0	63.0	98.0	39.0	60.0	97.0
Qwen2.5-7B	71.0	71.0	96.0	58.0	70.0	98.0
Baichuan2-7B	50.0	63.0	97.0	61.0	71.0	97.0
Llama-3.1	48.0	60.0	98.0	44.0	66.0	99.0
gemma-7b	50.0	55.0	93.0	51.0	58.0	94.0
Mistral-7B	52.0	63.0	99.0	52.0	57.0	96.0

Table 17: The accuracy of the baseline dataset composed of 100 words, which have the same form and the same meaning in Japanese and Chinese. "WM" stands for "Word Meaning", "WMC" for "Word Meaning Context", and "T" for "Translation".

set for which the LLM commits a homograph shortcut at least once out of the k translation samples. In contrast, Table 19 and 20 show the number of occurrences of the cross-lingual homographs themselves in translations. We let models generate translation for k times ($k = 1, 3, 5, 10$), and the total number of occurrence of the cross-lingual homographs are counted.

M Results after Removing Line Breaks

In the JKVC dataset, some lexical entries include unintended line breaks at the end of the definition field, and such artifacts were partially retained in our dataset. After removing these line breaks and rerunning the experiments, we observed changes in perplexity, with slight performance improvements for each model as shown in Table 21 and 22. Nevertheless, the overall performance trends remained consistent with the discussion presented in the main part of the paper. Therefore, we report only that this correction resulted in minor accuracy gains.

human		zh→ja	
		wrong	correct
ja→zh	wrong	21	12
	correct	7	578

llm-jp		zh→ja	
		wrong	correct
ja→zh	wrong	84	117
	correct	168	249

Elyza		zh→ja	
		wrong	correct
ja→zh	wrong	107	127
	correct	174	210

Qwen		zh→ja	
		wrong	correct
ja→zh	wrong	70	142
	correct	107	299

Baichuan		zh→ja	
		wrong	correct
ja→zh	wrong	96	154
	correct	127	241

Llama		zh→ja	
		wrong	correct
ja→zh	wrong	52	131
	correct	115	320

Gemma		zh→ja	
		wrong	correct
ja→zh	wrong	105	130
	correct	126	257

Mistral		zh→ja	
		wrong	correct
ja→zh	wrong	180	176
	correct	117	145

Table 18: Confusion matrices for bidirectional translation accuracy across 7 models

Model	$k = 1$			$k = 3$			$k = 5$			$k = 10$		
	T-1	T-2	T-4	T-1	T-2	T-4	T-1	T-2	T-4	T-1	T-2	T-4
llm-jp	111	34	65	326	116	188	534	195	316	1,097	388	637
ELYZA-JP	153	64	87	540	215	294	914	384	519	1,831	799	1,052
Qwen2.5-7B	119	60	72	401	175	244	669	287	432	1,363	585	863
Baichuan2-7B	185	73	104	552	219	312	922	365	520	1,847	730	1,040
Llama-3.1	130	60	84	395	175	246	663	281	413	1,327	553	794
gemma-7b	137	63	98	411	189	294	685	315	490	1,370	630	980
Mistral-7B	129	51	86	397	164	273	685	288	445	1,384	573	900

Table 19: The number of items that include cross-lingual homographs in the translation text itself in the Chinese-to-Japanese translation task. The translation generations have done for k times each ($k = 1, 3, 5, 10$).

N Performance of Models with Various Parameter Size

All primary experiments in this study were conducted using models with 7–8B parameters. To examine whether the homograph shortcut also appears in models of different parameter sizes, we further evaluated models with smaller parameter sizes. The results are presented in Table 23 and 24.

These results indicate a decreasing tendency in the occurrence of homograph shortcuts as model size increases. Moreover, even state-of-the-art GPT-4.1 is not fully immune to this phenomenon.

Model	$k = 1$			$k = 3$			$k = 5$			$k = 10$		
	T-1	T-2	T-3	T-1	T-2	T-3	T-1	T-2	T-3	T-1	T-2	T-3
llm-jp	125	69	123	345	221	356	579	351	591	1,139	712	1,163
ELYZA-JP	243	98	201	762	329	620	1,291	566	1,034	2,594	1,122	2,083
Qwen2.5-7B	80	55	94	247	167	265	422	275	436	841	560	863
Baichuan2-7B	316	133	251	947	397	753	1,579	663	1,255	3,159	1,328	2,510
Llama-3.1	128	65	119	370	194	348	607	321	554	1,214	662	1,100
gemma-7b	93	55	101	279	165	303	465	275	505	930	550	1,010
Mistral-7B	129	72	111	393	223	347	660	374	579	1,333	744	1,148

Table 20: The number of items that include cross-lingual homographs in the translation text itself in the Japanese-to-Chinese translation task. The translation generations have done for k times each ($k = 1, 3, 5, 10$).

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-3	T-1	T-2	T-3
Human	95.04	93.40	93.40	94.47	92.20	94.59	95.03	90.94
llm-jp	**76.94**	**37.36**	**79.00**	49.17	**58.36**	59.31	61.33	51.56
Llama-3	68.10	31.32	74.24	**51.38**	55.81	53.25	61.88	50.42
Qwen2.5-7B	55.39	33.52	59.09	44.20	27.20	**68.40**	71.27	59.77
Baichuan2-7B	43.97	26.37	58.87	44.20	46.74	61.69	65.75	48.73
Llama-3.1	53.88	29.12	67.75	45.30	46.74	71.43	**75.14**	**66.86**
gemma-7b	57.76	32.42	64.94	46.41	46.46	60.39	62.98	54.11
Mistral-7B	43.97	25.27	51.30	37.02	39.09	50.43	54.14	42.78

Table 21: Performance of LLMs on the Japanese word meaning task, word meaning in context task and Japanese-to-Chinese translation task without line breaks in the dataset.

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-4	T-1	T-2	T-4
Human	94.40	90.66	96.10	92.73	98.84	95.28	95.82	93.82
llm-jp	17.67	14.84	40.66	30.17	36.36	65.49	**72.63**	73.08
Llama-3	16.81	17.03	47.47	35.75	50.70	60.44	67.60	70.63
Qwen2.5-7B	45.91	26.37	72.97	48.04	53.85	64.18	70.39	**73.78**
Baichuan2-7B	**50.43**	26.37	**71.21**	**54.19**	**66.78**	58.90	62.57	64.34
Llama-3.1	31.03	19.78	63.96	46.93	45.45	**70.55**	71.51	72.38
gemma-7b	39.22	**28.02**	56.70	39.66	47.20	60.00	67.04	68.53
Mistral-7B	32.97	21.43	53.85	42.46	47.90	41.32	46.37	45.80

Table 22: Performance of LLMs on the Chinese word meaning task, word meaning in context task and Chinese-to-Japanese translation task without line breaks in the dataset.

This suggests that the proposed task remains a valid and meaningful challenge for current large language models.

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-3	T-1	T-2	T-3
Human	95.04	93.40	93.40	94.47	92.20	94.59	95.03	90.94
llm-jp-3-1.8b[11]	69.61	37.91	73.16	45.86	51.56	47.19	48.62	46.18
llm-jp-3-7.2b	76.94	37.36	79.00	49.17	58.36	59.31	61.33	51.56
llm-jp-3-13b[12]	76.72	39.56	79.44	50.83	52.69	61.69	63.54	54.39
Qwen2.5-1.5B[13]	37.72	26.37	42.64	37.57	23.23	63.64	67.96	55.24
Qwen2.5-3B[14]	50.86	29.67	53.90	39.78	30.88	69.26	68.51	62.32
Qwen2.5-7B	55.39	33.52	59.09	44.20	27.20	68.40	71.27	59.77
Qwen2.5-14B[15]	60.34	35.71	64.07	47.51	43.91	74.03	80.11	72.80
GPT-4.1	94.07	80.41	-	-	-	86.96	91.67	89.60

Table 23: Performance of LLMs of various parameters on the Japanese word meaning task, word meaning in context task and Japanese-to-Chinese translation task.

Model	Word Meaning		Word Meaning Context			Translation		
	T-1	T-2	T-1	T-2	T-4	T-1	T-2	T-4
Human	94.40	90.66	96.10	92.73	98.84	95.28	95.82	93.82
llm-jp-3-1.8b	19.83	14.29	33.41	25.70	37.76	60.44	65.36	68.53
llm-jp-3-7.2b	17.67	14.84	40.66	30.17	36.36	65.49	72.63	73.08
llm-jp-3-13b	16.38	12.64	37.14	27.93	40.91	65.27	73.74	76.22
Qwen2.5-1.5B	51.51	25.82	73.63	50.28	40.21	55.60	63.69	64.34
Qwen2.5-3B	48.06	26.92	70.77	45.81	49.65	58.24	65.36	67.48
Qwen2.5-7B	45.91	26.37	72.97	48.04	53.85	64.18	70.39	73.78
Qwen2.5-14B	45.91	28.57	74.73	53.07	50.70	71.21	75.98	80.77
GPT-4.1	66.65	67.99	-	-	-	85.82	90.26	83.65

Table 24: Performance of LLMs of various parameters on the Chinese word meaning task, word meaning in context task and Chinese-to-Japanese translation task.

[11] https://huggingface.co/llm-jp/llm-jp-3-1.8b-instruct3

[12] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct3

[13] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

[14] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

[15] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct