LLM-based Literal Example Generation for Japanese Multiword Expressions

Mio Ohashi, Hajime Kiyama, Zhidong Ling, Mamoru Komachi
Hitotsubashi University
{ohashi, hajime,shito,komachi}@scl.sds.hit-u.ac.jp

Abstract

We investigate whether GPT-5 can generate literal usage examples for Japanese multiword expressions (MWEs), whose literal readings are structurally low-frequency in available corpora. Prior work on MWEs has largely focused on detecting idiomatic usages in context, leaving literal usages underrepresented particularly for Japanese MWEs whose literal readings are rare and structurally diverse. Because literal readings are rarely attested in corpora, we design a lexicon-grounded setup that uses corpus non-literal usages as contrastive cues for controlled prompting. We evaluate the generated sentences using automatic literalness judgments and human literalness judgments, together with manual inspection. Our results show that providing contrastive non-literal information stabilizes literal generation and improves quality compared with prompts that include only literal information or no hints. The results indicate that Gemini-2.5-pro aligns more closely with human judgments for idiomatic interpretations than for literal ones, highlighting the relative difficulty of modeling literal readings of MWEs. The study demonstrates that GPT-5 can complement existing resources by supplying frequency-independent literal examples and offers a controlled framework for examining contextual meaning understanding of Japanese MWEs.

1 Introduction

Multiword expressions (MWEs) often show context-dependent meanings, posing a long-standing problem in natural language processing (Baldwin and Kim, 2010) A single expression may allow both literal and non-literal interpretations despite having an identical surface form, and many MWEs are semantically or syntactically non-compositional, with interpretations that vary depending on context (Schneider et al., 2014). For example, the English MWE *spill the beans* can refer either to physically dropping beans or idiomatically to revealing a secret.

Resource	Sentence Examples	lit. vs. non-lit.	Coverage
OpenMWE	✓	✓	Limited
JMWEL	✗	✗	✓
Ours	✓	✓	✓

Table 1: Comparison of existing Japanese MWE resources with respect to (i) availability of sentence examples, (ii) literal–idiomatic contrast, and (iii) coverage.

One key source of difficulty is the skewed distribution of literal and non-literal usages in corpora, where idiomatic uses dominate. For instance, the VNC-Tokens dataset (Cook et al., 2008), which annotates verb-noun combinations from the British National Corpus as idiomatic, literal, or unknown, contains 2,984 annotated tokens, with idiomatic usages substantially outnumbering literal ones. This imbalance is often more pronounced at the level of individual expressions: *get the sack* occurs far more frequently in its idiomatic sense than its literal one (43 vs. 7 tokens), while expressions such as *bring luck* appear almost exclusively idiomatically, with no literal attestations in the corpus. Because literal readings reflect compositional semantics, their correct modeling is essential for assessing whether models truly understand MWE meaning, rather than relying on frequency-based heuristics or memorized idiomatic patterns.

In this work, we focus on Japanese MWEs, where skewed usage distributions pose particularly severe challenges due to data scarcity and language-specific properties. Compared to English, the amount of freely reusable Japanese text data is substantially smaller, and typological properties of Japanese further exacerbate this imbalance. In particular, relatively free word order allows the components of a literal MWE to appear non-contiguously, complicating surface-based identification, whereas

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 307–325

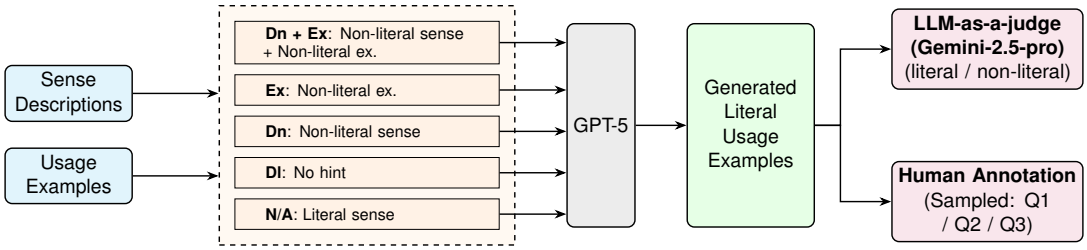

Figure 1: Overview of our framework for literal example generation for Japanese multiword expressions. The framework consists of controlled prompting under five conditions (Dn + Ex, Ex, Dn, Dl, N/A) that differ in the availability of dictionary definitions and non-literal usage examples, followed by automatic and human evaluation.

idiomatic usages tend to occur in more fixed forms (Hashimoto and Kawahara, 2008). For example, the expression *te-o nobasu* (literally, "extend one's hand") may appear discontinuously in its literal, physical sense, but occurs in a more compact form in its idiomatic meanings ("reach out," "expand into"). Because model evaluation relies on annotated examples, such skew directly constrains how MWE understanding can be assessed, making it difficult to construct resources that balance literal and idiomatic usages. Consequently, strong performance on MWE classification tasks may reflect frequency-based heuristics rather than true contextual understanding. Moreover, literal usages are not only rarer but also harder to detect, and the resulting imbalance is directly reflected in corpus statistics rather than being masked by data scale.

To address this gap, we investigate whether GPT-5 can be used to systematically generate literal usages of Japanese MWEs under controlled prompting conditions. Figure 1 provides an overview of our framework for literal example generation and evaluation. We target MWEs that appear both in JMWEL (Shudo et al., 2011) and JParaCrawl (Morishita et al., 2020). JMWEL is a structured lexicon of Japanese MWEs that provides linguistic annotations such as morphological segmentation, word-formation information, and syntactic patterns, and JParaCrawl is a large-scale Japanese–English parallel corpus constructed from web-crawled data, and evaluate the generated examples using automatic judgments and human judgments.

Our contributions are twofold. First, we propose a controlled framework for generating literal usage examples for Japanese MWEs and show that contrastive non-literal information improves generation stability. Second, through LLM-based analysis using Gemini-2.5-pro, we reveal that it captures idiomatic meanings more reliably than literal ones, highlighting a fundamental limitation in modeling compositional meaning. Together, these findings provide new insights into the capabilities and limitations of tested models in generating and evaluating literal meaning in MWEs.

Our main contributions are summarized as follows:

- We enable systematic analysis of literal usages of MWEs by generating sentence-level examples via controlled GPT-5 prompting, revealing the role of ambiguity in their interpretation.
- We apply this method to Japanese MWEs and show that the generated examples are often interpretable as literal usages, as confirmed by both LLM-based and human evaluation using .
- We conduct an LLM-based understanding test using Gemini-2.5-pro on generated examples intended as literal usages and show, by comparison with human judgments, that models capture idiomatic interpretations more reliably than literal ones.
- We construct a sentence-level resource contrasting literal and non-literal Japanese MWE usages, enabling controlled analysis of literal interpretation and revealing challenges in modeling literal meaning.

2 Related Work

MWE Resources OpenMWE (Hashimoto and Kawahara, 2009) contains Japanese MWEs extracted from corpora, and each example is annotated with labels such as *literal* and *idiomatic*, but its coverage is limited and the distributions of examples are often skewed, depending on how the MWEs happen to appear in the source corpora.

JMWEL (Shudo et al., 2018) provides detailed linguistic annotation, including morphological segmentation, word-formation information, and syntactic patterns, but lacks paired literal and non-literal usage examples, and context is largely absent, limiting its applicability to context-based NLP tasks.

In contrast, for English, several annotated MWE corpora have been constructed, including DiMSUM (Schneider et al., 2016), MAGPIE (Haagsma et al., 2020) and CoAM (Ide et al., 2025). However, comparable resources that systematically provide sentence-level contrasts between literal and non-literal usages remain limited for Japanese. Overall, existing MWE resources do not sufficiently support systematic and contrastive analysis of literal and non-literal usages, motivating the need for alternative approaches.

LLM-based Example Generation using GPT-5 Recently, LLMs have been used to generate contextual usage examples conditioned on dictionary information. Cassotti and Tahmasebi (2025) showed that English usage examples generated based on dictionary senses can achieve high quality and can be used as sense-specific data for diachronic semantic change research.

However, this line of research has not been directly applied to MWEs, which are often non-compositional and exhibit strong interactions between components. Compared to single words, MWEs pose different challenges, such as idiomatic readings dominating corpus examples and more complex syntactic behavior. Our work extends GPT-5-based usage generation to Japanese MWEs and analyzes how the model behaves when asked to generate literal usages for expressions whose non-literal reading is dominant in existing corpora.

3 Method

Task Definition and Framework We define a task in which GPT-5 generates literal usage examples for Japanese MWEs. Given a target MWE, GPT-5 is prompted, under conditions with different types of input information, to generate sentences in which the expression is used in its literal sense. Our framework consists of two stages: example generation and evaluation. Figure 1 illustrates the overall framework of our study. The aim of this setup is not to evaluate the performance of a particular model, but to clarify what kinds of information are effective for literal example generation.

Condition	Dictionary (Non-literal)	Dictionary (Literal)	Usage (Non-literal)
Dn + Ex	✓	–	✓
Ex	–	–	✓
Dn	✓	–	–
Dl	–	✓	–
N/A	–	–	–

Table 2: Information provided under each prompting condition. "Dictionary" refers to sense descriptions (literal or non-literal), and "Usage" refers to non-literal corpus examples.

Prompting Conditions In the generation stage, in addition to the target MWE, lexical information based on external resources such as dictionary descriptions and other MWE-related annotations is provided to GPT-5 in different combinations as prompting conditions. These prompting conditions are designed by controlling the availability of semantic cues related to literal and non-literal meanings. Specifically, we vary whether the model is provided with literal dictionary definition, non-literal dictionary definitions, and non-literal example usages. This design enables us to examine how different types of information influence the generation of literal usages, and whether contrastive signals between literal and non-literal meanings help suppress idiomatic interpretations.

We define five conditions (Dn + Ex, Ex, Dn, Dl, N/A) that differ in the combinations of input information, reflecting design choices regarding what information is provided to GPT-5. Dn + Ex provides both a non-literal usage example and a non-literal dictionary definition, whereas Ex and Dn provide only one of these two sources of non-literal information. Dl provides a literal dictionary definition. By contrast, N/A is a surface-form-only condition, in which no usage example or dictionary definition is given. For example, for the Japanese MWE *te o hiku*, the conditions differ in whether the model receives both a non-literal usage example and a non-literal dictionary definition, a non-literal usage example, a non-literal dictionary definition, a literal dictionary definition, or the expression alone.

While keeping the generation task itself fixed, we systematically vary the types of information supplied to the model, and the comparison of these conditions constitutes the core of our framework. Table 2 summarizes the prompting conditions used in this study.

Generation Setup We use GPT-5 as provided by the OpenAI API (gpt-5-2025-08-07) as the generation model and ask it to output five example sentences that use the expression in a literal way for each target MWE. The target set of MWEs consists of those that are listed in JMWEL and also appear in JParaCrawl. This design allows us to obtain both dictionary information and corpus-based reference examples.

For dictionary information, we use Nihon Kokugo Daijiten second edition (NKD) [1], a comprehensive dictionary of the Japanese language. We automatically separate the literal and non-literal senses of each target MWE based on the structure of the dictionary entries, which are available for 449 candidate expressions. Specifically, we use a heuristic based on lexical cues in the NKD definitions. Definition segments following phrases such as *no yoo ni* ('like/as if') and *koto kara* ('from the fact that') are treated as descriptions of non-literal or extended usages, whereas descriptions that preserve the compositional meanings of the component words are treated as literal senses. Depending on the prompt condition, we include or exclude these sense descriptions when constructing the input to GPT-5. We automatically separate the literal and non-literal senses of each target MWE based on the dictionary entries, which are available for 449 candidate expressions. Depending on the prompt condition, we include or exclude these sense descriptions when constructing the input to GPT-5.

For usage examples, we extract sentences containing the target MWEs from JParaCrawl using JMWEL as an index. Each extracted sentence is manually annotated as either a literal or a non-literal usage. The annotations were performed by two native Japanese speakers: the first author and a Japanese graduate student. Both annotators followed detailed annotation guidelines and annotated the data independently.

To examine how different types of lexical information provided in the prompt affect literal usage generation, we construct five prompting conditions. For each MWE and each condition, we ask GPT-5 to generate five sentences. The five prompting conditions (Dn + Ex, Ex, Dn, Dl, N/A) with a concrete example, showing what type of information is provided to the model in each condition are shown in the table in Appendix B. Table 3 summarizes the number of target MWEs and generated sentences for each prompting condition.

[1] https://japanknowledge.com/en/contents/nikkoku/index.html

	MWEs	Req.	Gen.
Dn + Ex	338	1,690	1,283
Ex	338	1,690	1,323
Dn	449	2,245	1,855
Dl	449	2,245	1,389
N/A	358	1,790	1,624

Table 3: Statistics of target MWEs and generated sentences for each prompting condition. Condition labels follow the definitions in Table 2. "Req." indicates the number of generation requests issued to the model, and "Gen." indicates the number of sentences successfully generated in the required format.

	Literal ratio	Cohen's κ
Dn + Ex	0.76	0.54
Ex	0.86	0.52
Dn	0.78	0.60
Dl	0.70	0.53
N/A	0.72	0.51
Overall	**0.76**	**0.54**

Table 4: Comparison across prompting conditions: human agreement literal ratio (treating non-literal and error as non-literal), inter-annotator agreement (Cohen's κ).

4 Evaluation

We employ human annotation to evaluate the quality of generated examples, focusing on their acceptability, literalness, and interpretability. Due to the cost and time required for manual annotation, we conduct human evaluation on a stratified sample of 250 sentences, consisting of 50 sentences per prompting condition. The evaluation set was obtained via stratified sampling from five generation conditions and was designed to represent the entire corpus to be released. Each MWE was treated as an independent instance with no duplication. For each MWE, we collected three types of judgments.

Q1: Literal Usage Availability (4-point scale). Annotators were asked to rate how easily they could imagine a literal usage of the expression, on a scale from 1 (Primarily non-literal) to 4 (Primarily literal).

Q2: Usage Category Classification. Annotators selected one of four categories describing the usage in context: *literal*, *non-literal*, *meaningless*, or *ambiguous*.

Q3: Interpretation Plausibility (5-point scale).

Q3 evaluates the plausibility of the generated sentence under two different interpretations. In Q3-1, annotators rated the plausibility of interpreting the sentence as *literal*, on a scale from 1 (completely unnatural) to 5 (very natural). In Q3-2, annotators rated the plausibility of interpreting the same sentence as *non-literal*, using the same 1–5 scale.

The annotation was conducted on Lancers, a Japanese crowdsourcing platform. Seven annotators participated in the study, all of whom were native speakers of Japanese.

The annotation was carried out in two stages. In the first stage, we conducted a pilot annotation for 100 sentences with three annotators. To make the pilot annotations comparable with the rest of the evaluation set, we then collected two additional independent annotations for the same 100 sentences. In the second stage, the remaining 150 sentences were annotated by five independent annotators. As a result, every sentence in the evaluation set received five independent judgments, while the total number of unique annotators was seven.

All evaluations were conducted individually without discussion among annotators.

Evaluation Metrics Q1–Q2: We report the overall distribution of votes across all instances (250 × 5 annotations) and present the proportion of each category in a table.

Q3: These ratings were treated as ordinal data, and the median of the five annotators was used as the representative value for each instance. This evaluation is intended not to measure the quality of a single generation condition, but to provide an overall human assessment of the entire released corpus. Detailed distributions and aggregated results are shown in the corresponding tables and figures.

5 Evaluation Results

Overall Comparison Across Conditions. We report two aspects of the generation results: human-annotated literal ratios and inter-annotator agreement in Table 4. The example-only condition achieves the highest literal ratio (0.86), while the dictionary-only condition yields the highest agreement (κ = 0.60). The no-hint setting performs worst on both metrics, suggesting a trade-off between literalness and annotator agreement. Overall, these results show that contrastive non-literal information is particularly effective for stabilizing literal generation.

Q1		Q2	
Primarily non-literal	33.8	Literal	72.1
Possibly literal	24.9	Non-literal	17.9
Depend on context	35.9	Meaningless	3.9
Primarily literal	5.4	Ambiguous	6.1

Table 5: Distribution of Q1 and Q2 Responses (All Votes, n = 1,250).

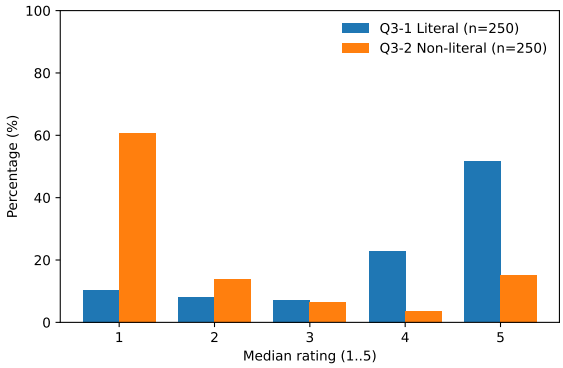

Figure 2: Distribution of Q3-1 (literal interpretation plausibility) and Q3-2 (non-literal interpretation plausibility). Each bar represents the median rating across five annotators.

Human Annotation Results. The overall distribution of Q1 and Q2 votes is shown in Table 5. Overall, *neutral* (35.9%) and *primarily non-literal* (33.8%) account for the largest proportions, suggesting that many MWEs are not clearly restricted to literal usage. In contrast, only 5.4% of instances were judged as *primarily literal*, indicating that expressions predominantly associated with literal usage are relatively rare.

A majority of the annotations (72.1%) were labeled as *literal*, which is consistent with the fact that the generated data were primarily designed to instantiate literal usages. At the same time, non-negligible proportions of *non-literal* (17.9%), *ambiguous* (6.1%), and *meaningless* (3.9%) labels were observed, indicating that some generated instances exhibit interpretive variability or perceived unnaturalness.

Interpretation Plausibility and Ambiguity. Figure 2 shows the distributions of Q3-1 and Q3-2. Q3-1 (literal interpretation plausibility) is concentrated at median values 4 and 5, indicating that literal interpretations are generally judged to be natural. In contrast, Q3-2 (non-literal interpretation plausibility) is heavily concentrated at median value 1, suggesting that non-literal interpretations are typically judged to be unnatural. These results

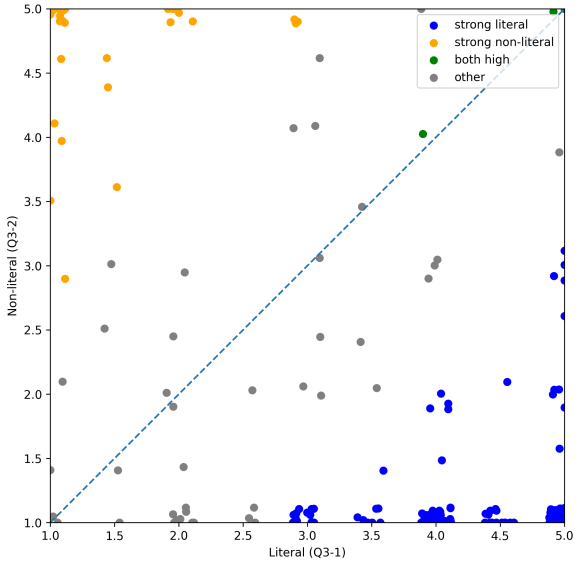

Figure 3: Item-level distribution of MWEs based on median naturalness ratings. Each point represents one MWE. The horizontal and vertical axes correspond to the naturalness of the literal and idiomatic interpretations, respectively. Points toward the lower right indicate literal preference, while points near the diagonal indicate interpretive ambiguity.

are consistent with the fact that all target data were generated with an intended literal interpretation. Nevertheless, a non-negligible number of instances receive median values of 4 or 5 in Q3-2.

To provide an item-level view of disagreement cases, we plot each MWE in a two-dimensional space defined by the median naturalness ratings for the literal and idiomatic interpretations (Figure 3). The horizontal axis represents the naturalness of the literal interpretation, and the vertical axis represents that of the idiomatic interpretation. A small amount of random jitter is added to avoid overlap.

This visualization complements the aggregated results by showing how individual cases are distributed. Cases classified as *strong_literal* and *strong_nonliteral* occupy opposite regions of the space, while *both_high* cases cluster near the diagonal, indicating interpretive ambiguity.

Importantly, the plot reveals that vote distributions and naturalness-based preferences do not always align. Some cases with a 3–2 vote split appear near the diagonal, whereas some 2–2–1 cases are located away from it. This suggests that vote distributions alone do not fully capture the underlying interpretive structure.

Metric	Literal	Non-literal
Exact Match (%)	40.62	74.38
± 1 Match (%)	71.25	89.38
MAE	0.784	0.353
Pearson r	0.745	0.856
QWK	0.660	0.853

Table 6: Results of LLM MWE Recognition Test using Gemini-2.5-pro.

6 LLM-based MWE Understanding Test using Gemini-2.5-pro

To further analyze how generated examples are interpreted, we conduct an LLM-based understanding test using Gemini-2.5-pro that compares model predictions with human judgments. This analysis allows us to examine whether Gemini-2.5-pro captures literal and idiomatic interpretations differently, even when the input data are designed to be literal.

Setup. We compared the Q3 scores (1–5 scale) assigned by human annotators with the corresponding scores predicted by Gemini-2.5-pro, and measured agreement using multiple evaluation metrics. The evaluation was conducted separately for *literal* and *non-literal* interpretations. Note that the evaluation is conducted on data primarily generated to reflect literal usages, rather than on a balanced set of literal and non-literal examples.

Evaluation. To compare predictions of Gemini-2.5-pro with human judgments, we compute expected accuracy, defined as the proportion of human votes that match the label predicted by Gemini-2.5-pro for each MWE.

For comparison with predictions of Gemini-2.5-pro, we retain the full distribution of the five human labels for each MWE. Specifically, suppose the human label distribution for a given MWE is (1, 2, 2, 0), meaning that one annotator selected category 1, two selected category 2, two selected category 3, and none selected category 4. If Gemini-2.5-pro predicts category 2, the expected accuracy for that MWE is calculated as the proportion of annotators who selected the same label (2/5 = 0.4 in this example). The final score is obtained by averaging the expected accuracy over all MWEs.

Results. The main results are summarized in Table 6. Overall, the model demonstrates particularly high agreement and correlation in idiomatic interpretations, suggesting that its semantic understand-

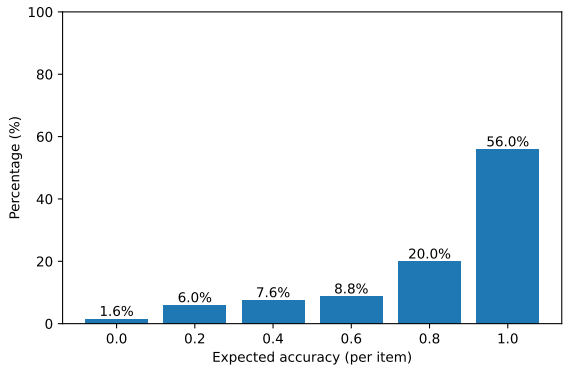

Figure 4: Expected accuracy distribution (Human vs Gemini-2.5-pro, n=250, macro=0.815).

ing of MWEs reaches a substantial level, although this result should be interpreted in light of the data distribution. Notably, agreement with human judgments is consistently higher for *non-literal* interpretations than for *literal* ones across all evaluation metrics. Since the generation task was originally designed to elicit *literal* usages, this gap may reflect the relative difficulty of producing and evaluating literal interpretations in this setting.

Figure 4 shows the distribution of expected accuracy between predictions of Gemini-2.5-pro and human judgments. The distribution is strongly skewed toward high agreement, with more than half of the instances achieving perfect matches between Gemini-2.5-pro and all annotators, and a large majority showing near-perfect agreement. This concentration suggests that the model's predictions align closely with human judgments in most cases. At the same time, the presence of a non-negligible tail at lower agreement levels indicates that certain instances remain sensitive to interpretative ambiguity, revealing limitations of the automatic classification in borderline cases.

Analysis of Disagreement Cases. To analyze disagreement cases, we examined the 15 instances with expected accuracy $= 0.2$ (i.e., 1/5 agreement). [2] Each instance is associated with a distribution of human votes over the four categories (c1–c4, corresponding to literal, non-literal, meaningless, and ambiguous), which directly reflects the degree of agreement or disagreement among annotators. Because each item has five annotators, this distribution results in expected accuracy values in increments of 0.2.

[2]The full list of these 15 instances is provided in Appendix C.

These cases fall into two types. This distinction helps separate cases where the model fails to capture a relatively stable human interpretation from those where the interpretation itself is ambiguous.

First, some instances show a clear human majority (4/5 votes), but Gemini-2.5-pro predicts a minority label. For example, *chikara o kasu* ("lend a hand") and *inochi o sasageru* ("devote one's life") were judged as *non-literal* by most annotators, while Gemini-2.5-pro predicted *ambiguous*. This suggests that the model may be influenced by the lexical abstractness of the expression rather than the specific context.

Second, in other instances, human judgments are already divided. Examples include *iki ga au* ("get along well"), *yume o miru* ("have a dream"). In such cases, both literal and non-literal interpretations remain plausible. Low expected accuracy therefore reflects interpretative variability rather than a clear model error.

7 Discussion

This section discusses the implications of our findings from both the human annotation and the LLM-based analysis using Gemini-2.5-pro. In particular, we focus on the interpretive structure of MWEs and its impact on literal usage generation.

7.1 Analysis of Human Judgments

This section uses human judgments to distinguish three types of disagreement cases in generated MWE examples. Because the target sentences were generated as literal examples, a disagreement case can have different implications for evaluation. Some cases are still naturally interpretable as literal usages. Some are more naturally interpreted idiomatically, and are therefore problematic as literal examples. Others are genuinely ambiguous, with both literal and idiomatic readings remaining plausible in context. Q2 provides categorical usage labels (*literal*, *non-literal*, *meaningless*, and *ambiguous*), but these labels alone do not show which of the above cases a split judgment represents. We therefore combine Q2 vote distributions with the Q3 naturalness ratings for literal and idiomatic interpretations. This allows us to analyze not only how annotators voted, but also which interpretation was judged more natural.

Interpretive Structure Behind Disagreement in Usage Judgments. In the human evaluation, annotators judged whether the usage of each MWE

in context was *literal* or *non-literal*. For some examples, however, annotators did not agree on the usage label. Instead, these disagreements reflect different types of interpretive structure.

In these cases, the usage label is divided, but one interpretation is rated as clearly more natural than the other. Both the literal and idiomatic interpretations are judged to be natural in the same context.

In the second type, the sentence allows both interpretations to coexist. This indicates a form of structural ambiguity.

Classification Method for Disagreement Cases. We quantify interpretive asymmetry using naturalness ratings.

Let $Q_{3,\mathrm{lit}}$ and $Q_{3,\mathrm{idi}}$ be the median naturalness scores for the literal and idiomatic interpretations, respectively. We define the difference between these values as

$$\Delta = Q_{3,\mathrm{lit}} - Q_{3,\mathrm{idi}}.$$

This value represents the relative naturalness of the two interpretations. A positive value of Δ indicates that the literal interpretation is more natural. Based on Δ, we classify disagreement cases into three types.

- **strong literal**: $\Delta \geq 2$
- **strong nonliteral**: $\Delta \leq -2$
- **both high**: $|\Delta| \leq 1$ and both interpretations have sufficiently high naturalness ($Q3 \geq 3$)

The *strong literal* and *strong nonliteral* categories represent cases where one interpretation is clearly preferred in the naturalness ratings. The *both high* category represents cases where both literal and idiomatic interpretations are relatively natural.

This classification allows us to distinguish two types of disagreement cases. In the first type, one interpretation is more stable. In the second type, both interpretations remain plausible in context.

Classification Results and Vote Distribution. Based on the criteria described in the previous section, disagreement cases were classified into three categories: *strong_literal*, *strong_nonliteral*, and *both_high*. The category *other* includes cases that do not meet the thresholds for these three main categories. Table 7 shows the distribution of these categories across the two vote patterns observed in Q2 (3–2 and 2–2–1).

	strong _literal	both_high	strong _nonliteral	other
3–2	5	2	1	7
2–2–1	4	5	3	8

Table 7: Distribution of disagreement categories across vote patterns (3–2 and 2–2–1).

Overall, the vote patterns show a certain correspondence with the naturalness-based classification. In the 3–2 pattern, asymmetric cases such as *strong_literal* and *strong_nonliteral* are relatively common. In contrast, the 2–2–1 pattern contains more *both_high* cases. This tendency reflects the difference between the two vote patterns: the 3–2 pattern indicates that a majority of annotators favor one interpretation, whereas the 2–2–1 pattern reflects a more evenly distributed judgment.

However, the vote distribution and the naturalness-based classification do not always align. For example, a 3–2 split suggests an asymmetric interpretation, yet some cases are classified as *both_high*. Conversely, the 2–2–1 pattern suggests structurally ambiguous cases, yet some examples are classified as *strong_literal* or *strong_nonliteral*.

These results suggest that the vote distribution alone does not fully capture the interpretive structure of disagreement cases. By incorporating naturalness judgments, we can better identify the underlying asymmetry or ambiguity in interpretation.

7.2 Implications for Generation Prompt Design

The analysis above has implications for prompt design in literal usage generation of MWEs. In this study, the task is to generate example sentences that illustrate the literal usage of an MWE. Therefore, the generated sentences are expected to be naturally interpretable under a literal reading. However, our analysis shows that some generated sentences are classified as *strong_nonliteral* in the naturalness evaluation. In these cases, the sentences are unnatural as literal examples, and the idiomatic interpretation is judged to be more natural.

In our prompt design, we explicitly instructed the model that it was not necessary to generate a sentence if a literal usage does not exist. Nevertheless, the overall generation rate of literal examples remained relatively high, and some of the generated sentences were classified as *strong_nonliteral*. This

result suggests that simply instructing the model to generate a literal example does not guarantee that the resulting sentence will be naturally interpretable under a literal reading. It also suggests that the model may prioritize producing a sentence over carefully determining whether a literal usage is actually available.

These observations highlight the importance of contextual specification in prompt design for literal usage generation. Rather than only instructing the model to produce a literal example, prompts should provide contextual information that supports the intended literal interpretation. This issue is particularly relevant for expressions whose idiomatic meanings are strongly conventionalized or for MWEs that allow both literal and idiomatic interpretations. For such expressions, the design of contextual information in the prompt can significantly affect how the generated sentence is interpreted. These results suggest that effective prompt design should provide contextual information that explicitly supports the intended literal interpretation, rather than relying solely on instructions.

8 Conclusion

This paper examined how GPT-5 generates literal usages of Japanese MWEs, focusing on the stability of generation under different prompting conditions. Our experiments showed that contrastive signals such as idiomatic (non-literal) sense definitions and non-literal usage examples play a crucial role in stabilizing literal example generation, yielding higher literalness and more consistent literalness judgments across conditions. Prompts with non-literal contrast produced more reliable literal usages, whereas the literal-definition-only condition was markedly less reliable, even for MWEs whose literal readings are rare. Overall, we propose a practical generation-based framework for supplementing literal examples of Japanese MWEs that are difficult to extract from corpora.

Limitations

This work has several limitations.

First, the applicability of the proposed framework is constrained by the availability and quality of language resources. For Japanese, we were able to leverage JMWEL as a structured MWE lexicon and JParaCrawl as a large-scale corpus, which enabled relatively fine-grained analyses of sense alternation and literal plausibility. However, the literal interpretability of MWEs is often gradient and depends on subtle constraints related to event conceptualization, verb–noun collocational preferences, and lexical semantic extension. For many other languages, comparable MWE lexicons and sufficiently large corpora are not available, making it difficult to operationalize these constraints at a similar level of granularity. As a result, while the framework is conceptually language-independent, its practical applicability is limited in low-resource settings.

Second, the proposed method partially relies on dictionary-based sense descriptions to distinguish literal and non-literal meanings and to guide generation. In this study, we use NKD as a representative dictionary resource to define sense boundaries and to provide explicit descriptions of literal meanings in the prompts. However, dictionary sense inventories and description styles vary across resources and languages, and they often abstract away from pragmatic and lexical constraints that affect actual usage. As shown by the qualitative analysis, even when a literal sense is theoretically available, its naturalness may be severely constrained. This reflects a broader mismatch between discrete dictionary sense representations and the gradient nature of literal plausibility in language use. In addition, our analysis suggests that GPT-5 and Gemini-2.5-pro may rely on superficial lexical or morphosyntactic cues when distinguishing literal from non-literal usage. For example, in Japanese MWEs such as *te o hiku*, the choice of case marker (e.g., *no te o hiku* vs. *kara te o hiku*) often correlates with the interpretation. This suggests that models may exploit such surface patterns rather than capture the underlying semantic distinction. Although GPT-5 can generate plausible literal usages under controlled conditions, this does not necessarily reflect robust semantic understanding, but rather sensitivity to distributional patterns only indirectly related to compositional meaning.

Third, our generation experiments use a single proprietary model, GPT-5. Therefore, the findings should be interpreted as a GPT-5-based case study rather than as evidence about LLMs in general. Although the controlled prompting framework itself is model-independent, the observed generation behavior may depend on the model family, training data, and model version. Future work should examine whether the same tendencies hold across other model families, including open-weight models and newer proprietary models.

Fourth, the evaluation methodology has inherent limitations. To ensure consistency across conditions, this study primarily relies on automatic judgments of literalness provided by Gemini-2.5-pro. While these judgments are generally reasonable, some generated sentences fall into borderline cases where the distinction between literal and non-literal interpretations is unclear even for human annotators. In particular, under the literal-definition-only condition, the model tends to produce pragmatically marked sentences or sentences with multiple plausible interpretations, which are often classified as non-literal by the evaluation model. These quantitative results should therefore be interpreted as reflecting relative tendencies across conditions rather than absolute measures of literal correctness. Future work includes calibrating these judgments with larger human-labeled samples and testing whether the same trends hold under alternative evaluation models.

Acknowledgments

This work was supported by the National Institute of Information and Communications Technology (NICT) under the "Research and Development of externally controllable modeling of multimodal information to enhance the accuracy of automatic translation."

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. Chapman and Hall/CRC.

Pierluigi Cassotti and Nina Tahmasebi. 2025. Sense-specific Historical Word Usage Generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001, Honolulu, Hawaii. Association for Computational Linguistics.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an Idiom Example Database for Supervised Idiom Identification. *Language Resources and Evaluation*, 43(4):355–384.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 27004–27021. Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A Comprehensive Dictionary of Multiword Expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 161–170, Portland, Oregon, USA. Association for Computational Linguistics.

Kosho Shudo, Toshifumi Tanabe, and Masahito Takahashi. 2018. Overview and Current Status of the Japanese Multiword Expression Lexicon JMWEL: Focusing on verbal multiword expressions. In *Proceedings of the 2018 Workshop on Language Resource Utilization*, pages 601–610. National Institute for Japanese Language and Linguistics.

A Generated Examples per Prompting Condition

Table 8 shows example for each prompting condition.

B Prompt Templates (Japanese Originals with English Translations)

Note: The experiments used the original Japanese prompts. For accessibility, we provide English

Linguistic information provided to the model (example with *te o hiku*)					
Dn + Ex	**Non-literal usage example and non-literal dictionary definition** *MWE*: te o hiku hand ACC pull Non-literal usage example: shijō kara te o hiku market from hand ACC pull Non-literal dictionary definition: lit. 'to withdraw from a relationship; to disengage'	Ex	**Non-literal usage example only** *MWE*: te o hiku hand ACC pull Non-literal usage example: shijō kara te o hiku market from hand ACC pull		
Dn	**Non-literal dictionary definition only** *MWE*: te o hiku hand ACC pull Non-literal dictionary definition: lit. 'to withdraw from a relationship; to disengage'	Dl	**Surface form only** *MWE*: te o hiku hand ACC pull	N/A	**Literal dictionary definition** *MWE*: te o hiku hand ACC pull Literal dictionary definition: lit. 'to pull someone by the hand; to lead'

Table 8: Prompting conditions (Dn + Ex, Ex, Dn, Dl, N/A). The table shows examples of the linguistic information provided to the model under each condition, using the Japanese MWE *te o hiku*. Romanization is shown for Japanese expressions; simplified morpheme-by-morpheme glosses annotate Japanese function words for the reader's convenience and were not included in the actual model inputs.

translations alongside the Japanese originals. All prompts instruct the model to output JSON only (without any additional explanation).

B.1 C1: Non-literal Dictionary + Non-literal Usage Example

B.1.1 SYSTEM prompt

Japanese (original)

あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、『文字通りの用法の日本語文』を作成するかどうかを判断し、必要なら文を生成します。
- 文字通りの用法が存在すると判断した場合は、その MWE を文字通りの意味で使った自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの用法が存在しない、または辞書上ほぼ比喩的・抽象的な意味に限られていると判断した場合でも、無理に文を作らなくて構いません。
- 入力として与えられる非文字通り用例（example_nonlit）と辞書定義（dict_def_nonlit）から MWE の意味を理解し、
文字通りの用法があるかどうかを慎重に判断し、has_literal に true/false を設定してください。
- 非慣用・非比喩。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は1つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

English (translation)

You are a Japanese sentence generation assistant.
- For each MWE (multiword expression / idiom), decide whether a literal usage exists and generate sentences only if needed.
- If a literal usage exists, generate up to {N_LIT} natural Japanese sentences using the MWE literally.
- If no literal usage exists, or if the dictionary meaning is almost entirely figurative/abstract, you do not need to force sentence generation.
- Understand the meaning of the MWE from

the provided non-literal usage example (example_nonlit) and the non-literal dictionary definition (dict_def_nonlit),
carefully judge whether a literal usage exists, and set has_literal to true/false.
- Sentences must be non-idiomatic and non-figurative; do not include simile markers such as “you ni” / “ka no you ni”.
- Each sentence must end with exactly one Japanese sentence-final period “。” (kuten).
- When generating multiple sentences for the same MWE, avoid redundancy in meaning and syntax as much as possible.
- Output JSON only, without any explanation.

B.1.2 USER prompt header

Japanese (original)
以下のデータに基づき、各 MWE について文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が存在すると判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文字通りの用法が存在しないと判断した場合は、has_literal を false、literal_list を空配列 [] にして構いません。
出力形式:

```
{
  "items": [
    {"mwe": "...", "has_literal": true,
"literal_list": ["...", "..."] },
    {"mwe": "...", "has_literal": false,
"literal_list": []},
    ...
  ]
}
```

データ:

English (translation).
Based on the following data, generate literal-usage Japanese sentences for each MWE.
- If you judge that a literal usage exists, return up to {N_LIT} sentences in literal_list.
- If you judge that no literal usage exists, you may set has_literal to false and literal_list to an empty array [].
Output format:

```
{
  "items": [
    {"mwe": "...", "has_literal": true,
"literal_list": ["...", "..."]},
    {"mwe": "...", "has_literal": false,
"literal_list": []},
    ...
  ]
}
```

Data:

B.2 C2: Non-literal Usage Example Only

B.2.1 SYSTEM prompt

Japanese (original)
あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、『文字通りの用法の日本語文』を作成するかどうかを判断し、必要なら文を生成します。
- 文字通りの用法が存在すると判断した場合は、その MWE を文字通りの意味で使った自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの用法が存在しない、または比喩的・抽象的な意味にほぼ限られていると判断した場合でも、無理に文を作らなくて構いません。
- 入力として与えられる非文字通り用例（example_nonlit）のみから MWE の意味を推測し、
文字通りの用法があるかどうかを慎重に判断し、has_literal に true/false を設定してください。
- 非慣用・非比喩。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は1つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

English (translation)
You are a Japanese sentence generation assistant.
- For each MWE, decide whether a literal usage exists and generate sentences only if needed.
- If a literal usage exists, generate up to {N_LIT} natural Japanese sentences using the MWE literally.
- If no literal usage exists, or if the meaning is almost entirely figurative/abstract, you do not need to force sentence generation.
- Infer the meaning of the MWE only from the provided non-literal usage example (example_nonlit),

carefully judge whether a literal usage exists, and set has_literal to true/false.
- Sentences must be non-idiomatic and non-figurative; do not include simile markers such as “you ni” / “ka no you ni”.
- Each sentence must end with exactly one Japanese sentence-final period “。”.
- Avoid redundancy across multiple sentences for the same MWE.
- Output JSON only, without any explanation.

B.2.2 USER prompt header

Japanese (original)
以下のデータに基づき、各 MWE について文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が存在すると判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文字通りの用法が存在しないと判断した場合は、has_literal を false、literal_list を空配列 [] にして構いません。
出力形式（C1と同様）とし、データを以下に与えます。
データ:

English (translation)
Based on the following data, generate literal-usage sentences for each MWE.
- If a literal usage exists, return up to {N_LIT} sentences in literal_list.
- If not, set has_literal to false and literal_list to [].
Use the same output schema as C1.
Data:

B.3 C3: Non-literal Dictionary Definition Only

B.3.1 SYSTEM prompt

Japanese (original)
あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、『文字通りの用法の日本語文』を作成するかどうかを判断し、必要なら文を生成します。
- 文字通りの用法が存在すると判断した場合は、その MWE を文字通りの意味で使った自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの用法が存在しない、または辞書上ほぼ比喩的・抽象的な意味に限られていると判断した場合でも、無理に文を作らなくて構いません。
- 入力として与えられる辞書定義（dict_mean_non_literal）から MWE の比喩的・慣用的な意味を理解し、
文字通りの用法があるかどうかを慎重に判断し、has_literal に true/false を設定してください。
- 非慣用・非比喩。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は1つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

English (translation)
You are a Japanese sentence generation assistant.
- For each MWE, decide whether a literal usage exists and generate sentences only if needed.
- If a literal usage exists, generate up to {N_LIT} natural Japanese sentences using the MWE literally.
- If no literal usage exists, or if the dictionary meaning is almost entirely figurative/abstract, you do not need to force sentence generation.
- Understand the figurative/idiomatic meaning from the non-literal dictionary definition (dict_mean_non_literal),
carefully judge whether a literal usage exists, and set has_literal to true/false.
- Sentences must be non-idiomatic and non-figurative; do not include simile markers such as “you ni” / “ka no you ni”.
- Each sentence must end with exactly one Japanese sentence-final period “。”.
- Avoid redundancy across multiple sentences for the same MWE.
- Output JSON only, without any explanation.

B.3.2 USER prompt header

Japanese (original)
以下のデータに基づき、各 MWE について文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が存在すると判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文字通りの用法が存在しないと判断した場合は、has_literal を false、literal_list を

空配列 [] にして構いません。
- 入力として与えられる情報は MWE 本体と非文字通りの辞書定義（dict_mean_non_literal）のみです。
出力形式（C1と同様）とし、データを以下に与えます。
データ:

English (translation)
Based on the following data, generate literal-usage sentences for each MWE.
- If a literal usage exists, return up to {N_LIT} sentences in literal_list.
- If not, set has_literal to false and literal_list to [].
- The input consists only of the MWE surface form and a non-literal dictionary definition (dict_mean_non_literal).
Use the same output schema as C1.
Data:

B.4 C4: No Hint (Surface Form Only)

B.4.1 SYSTEM prompt

Japanese (original)
あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、『文字通りの用法の日本語文』を作成するかどうかを判断し、必要なら文を生成します。
- 文字通りの用法が存在すると判断した場合は、その MWE を文字通りの意味で使った自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの用法が存在しない、または比喩的・抽象的な意味にほぼ限られていると判断した場合でも、無理に文を作らなくて構いません。
- 入力として与えられる情報は MWE の表層形だけです。辞書定義や非文字通り用例など、その他のヒントとなる情報は一切利用できないものと考えてください。
- 与えられた MWE の形だけから、文字通りの用法があるかどうかを慎重に判断し、has_literal に true/false を設定してください。
- 非慣用・非比喩。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は1つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

English (translation)
You are a Japanese sentence generation assistant.
- For each MWE, decide whether a literal usage exists and generate sentences only if needed.
- If a literal usage exists, generate up to {N_LIT} natural Japanese sentences using the MWE literally.
- If no literal usage exists, or if the meaning is almost entirely figurative/abstract, you do not need to force sentence generation.
- The only input available is the MWE surface form. Assume no dictionary definitions, no non-literal examples, and no other hints are available.
- Based only on the MWE form, carefully judge whether a literal usage exists and set has_literal to true/false.
- Sentences must be non-idiomatic and non-figurative; do not include simile markers such as “you ni” / “ka no you ni”.
- Each sentence must end with exactly one Japanese sentence-final period “。”.
- Avoid redundancy across multiple sentences for the same MWE.
- Output JSON only, without any explanation.

B.4.2 USER prompt header

Japanese (original)
以下の MWE の一覧について、文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が存在すると判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文字通りの用法が存在しないと判断した場合は、has_literal を false、literal_list を空配列 [] にして構いません。
出力形式（C1と同様）とし、データを以下に与えます。
データ:

English (translation)
For each MWE listed below, generate literal-usage Japanese sentences.
- If a literal usage exists, return up to {N_LIT} sentences in literal_list.
- If not, set has_literal to false and literal_list to [].
Use the same output schema as C1.
Data:

B.5 C5: Literal Dictionary Definition

B.5.1 SYSTEM prompt

Japanese (original)
あなたは日本語文作成支援ツールです。
- 各 MWE（慣用句）について、入力として与えられる『文字通りの意味の辞書定義（dict_mean_literal）』に基づき、
　その定義に合致する自然な日本語文を、最大 {N_LIT} 文まで作成してください。
- 文字通りの意味が定義されていても、文として成立させにくい場合は has_literal を false にしても構いませんが、
　定義がある場合は原則として作成を試みてください。
- 非慣用・非比喩的な文を作成すること。「ように／かのように」を含めてはいけない。
- 各文は句点「。」で終え、句点は1つのみとする。
- 同じ MWE について生成する複数の文は、できるだけ意味や構文が重複しないようにする。
- 出力は説明を含めず JSON のみ。

English (translation).
You are a Japanese sentence generation assistant.
- For each MWE, based on the provided literal dictionary definition (dict_mean_literal), generate natural Japanese sentences that match the definition, up to {N_LIT} sentences.
- Even if a Dl exists, if sentence generation is difficult, you may set has_literal to false; however, if a definition is given, you should generally attempt generation.
- Sentences must be non-idiomatic and non-figurative; do not include simile markers such as "you ni" / "ka no you ni".
- Each sentence must end with exactly one Japanese sentence-final period "。".
- Avoid redundancy across multiple sentences for the same MWE.
- Output JSON only, without any explanation.

B.5.2 USER prompt header

Japanese (original)
以下のデータに基づき、各 MWE について文字通りの用法の日本語文を生成してください。
- 各 MWE について、文字通りの用法が可能と判断した場合は literal_list に最大 {N_LIT} 文まで返してください。
- 文が作れないと判断した場合は、has_literal を false、literal_list を空配列 [] にして構いません。
- 入力として与えられる情報は MWE 本体と文字通りの辞書定義（dict_mean_literal）です。
出力形式:

```
{
  "items": [
    {"mwe": "...", "has_literal": true,
"literal_list": ["...", "..."] },
    {"mwe": "...", "has_literal": false,
"literal_list": []},
    ...
  ]
}
```

データ:

English (translation).
Based on the following data, generate literal-usage sentences for each MWE.
- If literal usage is possible, return up to {N_LIT} sentences in literal_list.
- If sentence generation is not feasible, set has_literal to false and literal_list to [].
- The input consists of the MWE surface form and a literal dictionary definition (dict_mean_literal).
Output format:

```
{
  "items": [
    {"mwe": "...", "has_literal": true,
"literal_list": ["...", "..."]},
    {"mwe": "...", "has_literal": false,
"literal_list": []},
    ...
  ]
}
```

Data:

C Instances with Expected Accuracy = 0.2

Table 9 lists all instances whose expected accuracy is 0.2. We report the human vote distribution as (c_1, c_2, c_3, c_4), corresponding to (*literal*, *non-literal*, *meaningless*, *ambiguous*), and the single label predicted by the Gemini-2.5-pro.

D Human Annotation Guidelines

This section presents the instructions provided to human annotators. The original instructions were written in Japanese. We provide an English translation below.

ID	MWE (romanized)	Human votes (c_1, c_2, c_3, c_4)	Gemini-2.5-pro pred.	ExpAcc
ex_000115	*kuchi o hiraku*	(4,0,0,1)	ambiguous	0.2
ex_000128	*tengu ni naru*	(4,1,0,0)	non-literal	0.2
ex_000140	*chikara o kasu*	(0,4,0,1)	ambiguous	0.2
ex_000144	*inochi o sasageru*	(0,4,0,1)	ambiguous	0.2
ex_000026	*na mo nai*	(4,0,0,1)	ambiguous	0.2
ex_000131	*iki ga au*	(1,3,0,1)	literal	0.2
ex_000142	*hana yori dango*	(3,1,0,1)	ambiguous	0.2
ex_000022	*you o tasu*	(1,1,1,2)	literal	0.2
ex_000042	*furidashi ni modoru*	(3,1,0,1)	ambiguous	0.2
ex_000048	*koshi o nukasu*	(1,1,2,1)	literal	0.2
ex_000054	*haisui no jin*	(1,1,1,2)	literal	0.2
ex_000062	*yume o miru*	(1,1,0,3)	non-literal	0.2
ex_000063	*te ga tsuku*	(2,1,1,1)	non-literal	0.2
ex_000064	*me ga hanasenai*	(1,3,0,1)	ambiguous	0.2
ex_000090	*kuchi o kiru*	(1,2,0,2)	literal	0.2

Table 9: All MWEs with expected accuracy $= 0.2$ (1/5 agreement).

Q1: Literal Usage Availability

For each multi-word expression (MWE) in the spreadsheet, please ignore the surrounding context and focus only on the expression itself. To what extent can you imagine a literal usage of this expression?

- 1: Basically idiomatic
- 2: Could be literal depending on the situation
- 3: Both literal and idiomatic usages are natural
- 4: Mainly used in a literal sense

Q2: Usage Category in Context

For each example sentence in the spreadsheet, how is the MWE used in that specific context?

- 1: Literal
- 2: Non-literal (idiomatic)
- 3: Meaningless / does not make sense
- 4: Ambiguous in this context

Q3-1: Plausibility under Literal Interpretation

When reading the sentence and interpreting the MWE as literal, how natural and acceptable is it in Japanese?

- 1: Not acceptable as a literal usage
- 2: Grammatically possible but very unnatural and rarely used
- 3: Somewhat unnatural but acceptable in limited contexts
- 4: Not very frequent but acceptable Japanese
- 5: Common usage and completely natural in Japanese

Q3-2: Plausibility under Non-literal Interpretation

When reading the sentence and interpreting the MWE as non-literal (idiomatic), how natural and acceptable is it in Japanese?

- 1: Not acceptable as an idiomatic usage
- 2: Grammatically possible but very unnatural and rarely used
- 3: Somewhat unnatural but acceptable in limited contexts
- 4: Not very frequent but acceptable Japanese
- 5: Common usage and completely natural in Japanese

E Gemini-2.5-pro Prompt

This section describes the prompt used in the LLM-based MWE understanding test using Gemini-2.5-pro. The original prompt was written in Japanese. An English translation is provided below for reproducibility.

System Instructions

The model was instructed as follows:

> Evaluate the MWE in each sentence under two independent interpretations:

A: Literal reading (interpret the expression at face value, without using idiomatic or metaphorical meaning)

B: Idiomatic reading (interpret the expression using its conventional idiomatic meaning)

Rate the naturalness on a 1–5 scale: 1 = Not acceptable 2 = Extremely unnatural 3 = Acceptable only in limited contexts 4 = Natural 5 = Very natural

For each item, output exactly one line of JSON. The keys and their order must be: id, mwe, literal_score, idiom_score. Do not output any explanations or additional text.

F Literal Usage Examples with Glosses

Note: This appendix presents only literal *usage examples. Original Japanese sentences are romanized. Minimal glosses annotate Japanese function words (e.g., case particles), followed by an English translation of the whole sentence.*

F.1 Target MWE: *te o hiku* ("pull someone's hand")

F.1.1 C1: Non-literal Dictionary + Non-literal Usage Example

(B1) *haha wa kodomo no te o hiite ōdan-hodō o watatta*
mother TOP child GEN hand ACC pull.GER crosswalk ACC cross.PST
'The mother pulled her child's hand and crossed the crosswalk.'

(B2) *annainin ga kurai dōkutsu de kyaku no te o hiita*
guide NOM dark cave LOC guest GEN hand ACC pull.PST
'The guide pulled the guest's hand in the dark cave.'

(B3) *kare wa ashimoto no obotsukanai sobo no te o hiite kaidan o agatta*
he TOP footing GEN unsteady grandmother GEN hand ACC pull.GER stairs ACC go.up.PST
'He pulled his unsteady grandmother's hand and went up the stairs.'

(B4) *kanojo wa korobikaketa yūjin no te o hiite taisei o tatenaosaseta*
she TOP stumble.about.to friend GEN hand ACC pull.GER posture ACC recover.CAUS.PST
'She pulled her friend's hand and helped them regain balance.'

(B5) *keisatsukan wa maigo no te o hiite kōban e tsureteitta*
police.officer TOP lost.child GEN hand ACC pull.GER police.box ALL take.PST
'The police officer pulled the lost child's hand' 'and took them to the police box.'

F.1.2 C2: Non-literal Usage Example Only

(B6) *hahaoya ga kodomo no te o hiite ōdan-hodō o watatta*
mother NOM child GEN hand ACC pull.GER crosswalk ACC cross.PST
'The mother pulled her child's hand and crossed the crosswalk.'

(B7) *keikan ga rōjin no te o hiite kaidan o nobotta*
police.officer NOM elderly.person GEN hand ACC pull.GER stairs ACC climb.PST
'The police officer pulled the elderly person's hand and climbed the stairs.'

(B8) *sofu ga mago no te o hiite yamamichi o aruita*
grandfather NOM grandchild GEN hand ACC pull.GER mountain.path ACC walk.PST
'The grandfather pulled his grandchild's hand and walked along the mountain path.'

(B9) *kanojo wa maigo no te o hiite annaijō e mukatta*
she TOP lost.child GEN hand ACC pull.GER information.desk ALL head.PST
'She pulled the lost child's hand and headed to the information desk.'

(B10) *hoikushi* *ga* *enji* *no*
nursery.teacher NOM child GEN
o *hiite* *retsu* *o*
ACC pull.GER line ACC
totonoeta
arrange.PST
'The nursery teacher pulled the child's hand and arranged the line.'

F.1.3 C3: Non-literal Dictionary Definition Only

(B11) *haha* *wa* *hitogomi* *no* *naka*
mother TOP crowd GEN inside
de *osanai* *ko* *no* *te*
LOC small child GEN hand
o *hiita*
ACC pull.PST
'The mother pulled her small child's hand in the crowd.'

(B12) *sensei* *ga* *enji* *no* *te*
teacher NOM child GEN hand
o *hiite* *ōdan-hodō* *o*
ACC pull.GER crosswalk ACC
watatta
cross.PST
'The teacher pulled the child's hand and crossed the crosswalk.'

(B13) *kyūjotaiin* *wa* *kurai* *tonneru*
rescuer TOP dark tunnel
de *fushōsha* *no te*
LOC injured.person GEN hand
o *hiita*
ACC pull.PST
'The rescuer pulled the injured person's hand in the dark tunnel.'

(B14) *sofu* *wa* *ishidan* *de*
grandfather TOP stone.steps LOC
mago *no* *te* *o*
grandchild GEN hand ACC
hiki, *yukkuri* *agatta*
pull.GER slowly go.up.PST
'The grandfather pulled his grandchild's hand and slowly went up the stone steps.'

(B15) *ani* *wa* *maigo* *ni*
older.brother TOP lost DAT
natta *otōto* *no*
become.PST younger.brother GEN
te *o* *hiite* *eki*
hand ACC pull.GER station
e *mukatta*
ALL head.PST
'The older brother pulled his lost younger brother's hand and headed to the station.'

F.1.4 C4: No Hint (Surface Form Only)

(B16) *hahaoya* *ga* *kodomo* *no* *te*
mother NOM child GEN hand
o hiku
ACC pull.PRS
'The mother pulls her child's hand.'

(B17) *kaigosha* *ga* *kōreisha* *no*
caregiver NOM elderly.person GEN
te *o* *hiku*
hand ACC pull.PRS
'The caregiver pulls the elderly person's hand.'

(B18) *eki* *no* *kaisatsu* *de* *chichioya*
station GEN gate LOC father
ga *musume* *no* *te* *o*
NOM daughter GEN hand ACC
hiku
pull.PRS
'At the ticket gate, the father pulls his daughter's hand.'

(B19) *kurai* *sakamichi* *de* *yūjin*
dark slope LOC friend
no *te* *o* *hiku*
GEN hand ACC pull.PRS
'On the dark slope, someone pulls their friend's hand.'

(B20) *hoikushi* *ga* *enji* *no*
nursery.teacher NOM child GEN
te *o* *hiku*
hand ACC pull.PRS
'The nursery teacher pulls the child's hand.'

F.1.5 C5: Literal Dictionary Definition

(B21) *hahaoya* *ga* *osanai* *ko* *no*
mother NOM small child GEN
te *o* *hiite* *ōdan-hodō*
hand ACC pull.GER crosswalk
o *watatta*
ACC cross.PST
'The mother pulled her small child's hand and crossed the crosswalk.'

(B22) *kare wa ashiba no warui*
he TOP footing bad
sakamichi de yūjin no te
slope LOC friend GEN hand
o hiite nobotta
ACC pull.GER climb.PST
'He pulled his friend's hand and climbed the' slope with poor footing.'

(B23) *gaido ga kankōkyaku no te*
guide NOM tourist GEN hand
o hiite dōkutsu no
ACC pull.GER cave GEN
naka o susunda
inside ACC proceed.PST
'The guide pulled the tourist's hand and proceeded inside the cave.'

(B24) *ani ga otōto*
older.brother NOM younger.brother
no te o hiite
GEN hand ACC pull.GER
eki no kaisatsu o
station GEN gate ACC
nuketa
pass.PST
'The older brother pulled his younger brother's hand and passed through the station gate.'

(B25) *futari wa tagai ni*
two.people TOP each.other DAT
te o hiite wa
hand ACC pull.GER circle
ni natte odotta
DAT become.GER dance.PST
'The two people pulled each other's hands and danced in a circle.'