Challenges in Japanese Euphemism Classification: An Analysis of Pretrained Japanese and Multilingual Models

Noriko Takahashi, Whitney Poh, Libby Barak, Jing Peng, Anna Feldman
Montclair State University
1 Normal Ave, Montclair, NJ 07043, USA
{takahashin1, pohw1, barakl, pengj, feldmana}@montclair.edu

Abstract

Euphemisms present a persistent challenge for NLP because their interpretation depends on pragmatic inference, social norms, and contextual cues rather than surface meaning alone. Although Potentially Euphemistic Terms (PET)-based resources have been developed for several languages, Japanese euphemisms remain computationally unexplored despite their close interaction with honorifics, register variation, and orthographic choice. We introduce **JP-PET**, the first PET-based dataset for Japanese euphemism classification, comprising 1,672 annotated sentences across 101 PETs and ten semantic domains with register metadata. We evaluate two Japanese monolingual transformer models (Rinna RoBERTa and Tohoku BERT) and the multilingual XLM-R under three controlled PET-level data splits that isolate lexical familiarity and generalization to unseen euphemisms. While models achieve strong performance when PETs are shared between training and test data, performance drops substantially under PET-disjoint conditions, indicating reliance on lexical familiarity. Error analysis suggests potential challenges in politically conventionalized expressions, metaphor-based euphemisms, and orthographic mitigation strategies. JP-PET provides the first benchmark for studying pragmatic meaning in Japanese NLP.

Keywords: Japanese NLP, euphemism detection, pragmatic inference, PETs, language models, register variation

1. Introduction

Natural Language Processing (NLP) continues to face challenges in interpreting indirect language uses, where meaning depends on pragmatic inference and cultural knowledge rather than literal expression. Recent work shows that despite substantial progress on various NLP tasks, pretrained transformer still struggle with the processing non-literal expressions, including figurative language that is novel in the sense of being rare, recently coined, or not frequently observed during pretraining, as well as culturally specific expressions (Liu et al., 2022; Jang et al., 2023; Ichien et al., 2024). Euphemisms pose a related challenge: they soften or reframe direct meanings to maintain politeness or avoid discomfort, so interpretation depends on social norms and contextual cues rather than surface semantics (Allan and Burridge, 1991; Pinker, 2003). Prior studies show that pretrained transformer models often struggle with euphemisms, whose interpretation depends on context, pragmatic inference, or distinguishing literal from euphemistic uses, as well as expressions involving metonymy and metaphor (Gavidia et al., 2022; Lee et al., 2024). These limitations highlight a central challenge for euphemism research: distinguishing between the literal and euphemistic uses of Potentially Euphemistic Terms (PETs). Many PETs retain their literal meaning in some contexts while becoming euphemistic in others, making them difficult for models to classify reliably. For example, the phrase “the bird and the bees” could refer to actual birds and bees as animals, but also refer to the sex-ed talk.

While recent work in English (Lee et al., 2024; Gavidia et al., 2022; Lee et al., 2023) has begun to address this challenge by developing PET frameworks and euphemism-detection benchmarks, these efforts remain almost entirely language-specific. Japanese euphemisms, in particular, remain largely unexplored, even though Japanese communication is strongly shaped by indirectness, contextual interpretation, and sensitivity to social relationships (Maynard, 1997; Cook, 2006). As pretrained transformer continue to be integrated into education, translation, writing assistance, and customer-facing computer applications, it is crucial to evaluate how well they handle these nuanced forms of indirect meaning. Understanding pretrained transformer performance in regards to Japanese euphemisms will not only highlight current limitations but will also inform the development of systems that must operate reliably in culturally sensitive contexts. This study, therefore, examines Japanese euphemistic expressions and evaluates how pretrained transformer interpret them, with the goal of supporting future applications that require accurate and contextually appropriate handling of Japanese pragmatic meaning.

To address the gap identified above, and to establish a foundation for computational work on Japanese euphemisms, this study makes three primary contributions.

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 1–11
11 May 2026.

- Introduces the first PET-based Japanese euphemism dataset, including metadata such as domain, register, and orthographic variation.
- Evaluates Japanese and multilingual pretrained models on distinguishing literal vs. euphemistic PET usage.
- Conducts a focused error analysis to identify key challenges in modeling Japanese euphemisms.

2. Potentially Euphemistic Terms (PETs)

To capture the ambiguity inherent in euphemistic expressions, NLP researchers introduced the concept of potentially euphemistic terms (PETs), which are words or phrases whose meaning can be literal or euphemistic depending on context (Lee et al., 2022; Gavidia et al., 2022). For instance, *between jobs* may literally describe a temporary career transition (e.g., *She is between jobs after finishing her contract*) or function euphemistically to mean unemployed (e.g., *He has been between jobs for over a year*) (Lee et al., 2022). Similarly, *special* may denote a general sense of uniqueness but can also serve as a euphemism for disability in expressions such as *special needs*. These examples illustrate how a single expression can function as a euphemism in some contexts but not in others. Because language and social norms evolve, and because interpretation varies across individuals, annotators often disagree on whether an expression should be labeled as euphemistic. PETs therefore offer a practical framework for annotation and modeling. As noted by Gavidia et al. (2022), corpora annotated for PETs capture this pragmatic variability and highlight the importance of context-sensitive interpretation. Modeling PETs requires distinguishing pragmatic meaning from surface semantics, which makes the task considerably more complex than standard lexical classification.

3. Linguistic Characteristics of Euphemisms

Euphemisms are expressions that allow speakers to avoid direct meaning, hide uncomfortable truths, or convey politeness through indirect speech in order to maintain social relationships (Rababah, 2014). As Allan and Burridge (1991) explain, euphemistic expressions soften the impact of a message and avoid taboo wording, and they commonly appear in culturally sensitive domains such as death, illness, sexuality, disability, aging, and social class (Casas Gómez, 2009; Valentine, 1998). Because euphemisms are tied to social norms, their meanings vary across communities and registers and change over time. Pinker (2003) describes this process as the ”euphemism treadmill,” in which expressions gradually lose their softening effect as they become associated with the taboo concepts they refer to, leading speakers to introduce new forms (Pinker, 2007).

Euphemisms are often realized through strategies such as circumlocution, lexical substitution, abbreviation, and semantic generalization (Allan and Burridge, 2006). While these mechanisms are broadly shared across languages, the specific expressions and their social motivations are language-specific, which makes their computational modeling highly dependent on cultural and contextual knowledge.

4. Euphemisms in Japanese

4.1. Domains of Japanese Euphemisms

Japanese euphemisms appear across common taboo domains such as *Death & Dying, Bodily Functions, Sexuality & Relationships, Crime & Social Issues,* and *Illness & Disability* (Allan and Burridge, 1991; Casas Gómez, 2009). They are also frequently used in domains such as Appearance, Personality, *Workplace & Economy*, and *Aging*. For example, 旅立つ (tabidatsu; literal meaning: to set out on a journey) is used euphemistically to mean ‘to die’ (Maynard, 1997), and ぽっちゃり (pochari) provides a Japanese mimetic expression that conveys a softened description of body size (Hamano, 1998).

Although these domains broadly overlap with those observed in English, the specific expressions and their sociocultural motivations differ, reflecting language-specific norms of politeness and indirectness.

4.2. Japanese Orthography

Japanese orthography allows a single lexical item to be written in kanji, hiragana, or katakana. These choices can signal differences in tone and social nuance. Orthographic variation therefore plays an important role in euphemistic expression by softening negative connotations or creating distance from sensitive meanings. For example, the word 障害 (shōgai) ‘disability’ is sometimes written as 障がい, where the kanji 害 (‘harm’) is replaced with the hiragana がい to reduce harshness. Another example involves katakana: the kanji 物 (mono/butsu) ‘thing’ can appear as ブツ (butsu), a slang form that can refer to drugs or contraband and thus carries a more implicit, coded meaning. By using katakana instead of kanji, the term de-

parts from the literal sense ‘thing’ and signals an indirect reference.

These variations introduce additional challenges for computational modeling, as identical pronunciations may correspond to different pragmatic functions depending on the script. Models must therefore recognize orthographic choices as signals of pragmatic intent rather than treating all forms as equivalent.

4.3. Register Variation

Register	PET	Literal	Meaning
Formal	お亡くなりになる (onakunari ni naru)	gone	to pass away
Neutral	女の子の日 (onnanoko no hi)	girl’s day	menstruation
Informal	ボンビー (bonbii)	–	‘poor’

Table 1: Examples of Japanese euphemistic expressions (PETs) across registers, with literal gloss and euphemistic meaning.

Register refers to the level of politeness and social distance conveyed through linguistic expression, and in Japanese, euphemisms often vary by register even when they express the same underlying meaning (Maynard, 1997). In this study, we group Japanese euphemisms into three levels: formal, neutral, and informal, reflecting differences in situational context and social expectations (Biber and Conrad, 2019). These variations create challenges for computational models, as the same meaning can be realized through forms that differ in politeness and social intent. Models must therefore rely on contextual cues beyond surface meaning to correctly interpret euphemistic usage. See Table 1 for an example for each level. The informal example ボンビー (bonbii) does not have a direct literal meaning, as it is a phonological modification of 貧乏（binbo-) ’poor’ used to soften or stylize the expression.

5. Research Questions

Japanese euphemisms challenge computational modeling because they vary across domains, permit orthographic alternations, and shift register with social context. Because PET-based resources do not capture these Japanese-specific patterns, we establish a baseline for Japanese euphemism classification using pretrained Japanese and multilingual models and analyze their generalization across controlled PET-level splits.

RQ1. How reliably do pretrained Japanese and multilingual language models classify euphemistic expressions across domains and registers?

RQ2. In which domains and registers do these models most frequently produce errors, and what insights do these patterns provide for future euphemism modeling and dataset construction?

Here, domains refer to semantic fields, and register refers to politeness level (formal, neutral, informal).

6. Methodology

This study frames euphemism detection as a binary classification task. For each sentence containing a potentially euphemistic term (PET), we insert [PET_BOUNDARY] markers around the target expression and provide the sentence to the model as input. The model predicts whether the marked PET is used euphemistically (1) or literally (0). We fine-tune three pretrained transformer models Rinna (rinna Co., Ltd., 2023), Tohoku (University, 2020), and XLM-Roberta (Conneau et al., 2020), on this task and evaluate their generalization across two PET-level data splits designed to test familiar-item learning and strict no-overlap generalization to new euphemisms.

6.1. PETs List and Data Collection

We collected data from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) via Shonagon, which covers diverse genres (e.g., news, magazines, books, web, official documents) and supports variation in register and orthography. We compiled 100 PETs based on prior PET work (Lee et al., 2022; Biyik et al., 2024), Japanese pragmatics and politeness research (Maynard, 1997; Cook, 2006), and native-speaker judgment, and queried BCCWJ to retrieve naturally occurring sentences containing inflected and orthographic variants.

6.2. Automatic PET Extraction

We used Sudachi (Takaoka et al., 2018) for morphological analysis and customized rules to capture inflectional and orthographic variants of PETs. Detected PET spans were marked with [PET_BOUNDARY] so the model could focus on the target region (e.g., Input: 「彼は去年、亡くなりました。」 ’He passed away last year.’ Annotated:「彼は去年、[PET_BOUNDARY] 亡くなりました [PET_BOUNDARY]。」‘He [PET_BOUNDARY] passed away [PET_BOUNDARY] last year.’).

6.3. Annotation Procedure

Three native Japanese speakers annotated each PET instance as euphemistic (1) or literal (0) us-

Model	Type	Tokenization	Pretraining Data	Coverage	Main Characteristics
Rinna	Monolingual	SentencePiece (subword)	Japanese CC-100 + Wikipedia	Japanese	Broad web coverage; captures modern usage and stylistic variation
Tohoku BERT	Monolingual	MeCab + UniDic → WordPiece	Japanese Wikipedia	Japanese	Linguistically informed segmentation; encyclopedia-style language
XLM-R	Multilingual	SentencePiece (subword)	Multilingual CC-100	100+ languages	Large-scale multilingual training; strong generalization

Table 2: Comparison of models, tokenization, and pretraining data.

ing shared guidelines with examples of clear and ambiguous cases. Inter-annotator agreement was high (Fleiss' κ = 0.87, 95% CI [0.85, 0.89]), and final labels were determined by majority vote.

6.4. Dataset Overview

JP-PET contains 1,672 annotated sentences for 101 PETs spanning multiple domains (Table 3) Per-PET statistics are reported in Appendix A.1.

Domain	Count	Percent (%)
Death & Dying	406	24.3
Bodily Functions	299	17.9
Sexuality & Relationships	224	13.4
Crime & Social Issues	168	10.1
Illness & Disability	146	8.7
Appearance	98	5.9
Personality	86	5.1
Workplace & Economy	83	5.0
Aging	69	4.1
Others	63	3.8
Politics & War	30	1.8
Total	1,672	100.0

Table 3: Distribution of domains in the JP-PET dataset.

Note: Domains correspond to semantic categories commonly observed in euphemism research (Allan and Burridge, 2006).

6.5. Dataset Splitting

To investigate how Japanese euphemisms vary in difficulty across semantic domains and individual Potentially Euphemistic Terms (PETs), we designed two complementary data-splitting schemes. Both splits use the same canonical test pool to ensure that model performance is directly comparable across conditions. In Split 1, PETs may appear in the training, validation, and test sets, allowing lexical overlap. In Split 2, PETs in the test set are strictly unseen during training, enforcing a no-overlap condition.

6.5.1. Always vs. sometimes euphemisms

For each PET, we first examined its label distribution. PETs annotated as euphemistic in all occurrences (label = 1) were categorized as always-euphemistic, whereas PETs that appeared with both euphemistic and non-euphemistic labels were categorized as sometimes-euphemistic. Because always-euphemistic expressions provide stable pragmatic meaning, all of their sentences were reserved for the test pool to probe model performance on consistently euphemistic items.

6.5.2. Canonical shared test pool and 10 folds

The canonical test pool consisted of:

- all sentences containing always-euphemistic PETs, and
- a stratified sample of sentences for sometimes-euphemistic PETs, balanced across labels (1 = euphemistic, 0 = non-euphemistic) for each PET.

This pool was partitioned into 10 disjoint folds, ensuring that across the 10 folds, every PET instance appears exactly once in a test set. This design enables fine-grained inspection of model performance on all PETs while keeping the evaluation consistent across both splits.

6.5.3. Split 1 – Shared PETs across train/validation/test

In Split 1, PETs are allowed to appear in all partitions. After constructing the canonical test pool, the remaining sometimes-euphemistic sentences were divided into training and validation sets using label-balanced sampling within each PET. The test sets correspond exactly to the predefined canonical folds; no additional test instances were created.

The test folds (Split 1-1 through 1-10) follow the fixed test partitions, while train and validation remain constant across folds. This split reflects a standard setting in which the model encounters the same PET string during training and testing but in

different contexts, allowing us to measure contextual generalization without lexical novelty.

6.5.4. Split 2 – PET-disjoint train/validation/test

Split 2 reuses the exact same 10 test folds as Split 1 but enforces complete PET disjointness across all partitions. No PET in the test set appears in train or validation, and no PET in validation appears in train. For each fold, entire PETs (not single sentences) were assigned to train or validation, while keeping label balance within each PET as much as possible. This split evaluates zero-shot PET generalization – whether a model can recognize euphemistic meaning for lexically unseen PETs.

6.5.5. Data Partitioning and Label Balancing

Across all splits, data are partitioned into training, validation, and test sets using an approximate 8:1:1 ratio, subject to PET-level constraints. For sometimes-euphemistic PETs, we perform label balancing by sampling approximately equal numbers of euphemistic (1) and non-euphemistic (0) instances within each PET and split. Always-euphemistic PETs appear exclusively in the test sets, which facilitates comparison of model behavior for stable euphemisms across Splits 1–2.

The distribution of test instances by domain and label is reported in Appendix A.2. Within the sometimes-euphemistic subset, the test data remain close to balanced across labels. Small imbalances occur for PETs with inherently uneven label distributions, where perfect balancing is not possible. Nevertheless, the overall distribution remains close to balanced and supports stable evaluation across domains.

The dataset is publicly available.[1]

6.6. Models

We fine-tuned three transformer models (Table 2): Rinna (monolingual RoBERTa; SentencePiece; (rinna Co., Ltd., 2023)), Tohoku (monolingual BERT; MeCab/UniDic + WordPiece; (University, 2020)), and XLM-R (multilingual; SentencePiece; (Conneau et al., 2020)). We used max length 320, batch size 16, learning rate 2×10^{-5}, up to 8 epochs with early stopping, and selected models by validation macro-F1. Results are averaged over five runs.

As a baseline, we train a linear classifier over a frozen encoder. Each input instance is encoded by a pretrained model, and the final-layer [CLS] representation is used as a fixed feature vector. A logistic regression classifier is trained on these [CLS] representations without updating the encoder. The baseline is trained and evaluated on the same data splits as the fine-tuned models.

[1] https://github.com/NLPlabMSU/jp-pet-dataset.git

	Split 1 (Seen)		Split 2 (Unseen)	
Model	Baseline	Fine-tuned	Baseline	Fine-tuned
Rinna RoBERTa	0.45	0.68	0.46	0.57
Tohoku BERT	0.50	**0.73**	0.51	0.59
XLM-RoBERTa	0.48	**0.73**	0.52	**0.61**

Table 4: Macro-F1 scores for baseline and fine-tuned models across Split 1 (seen expressions) and Split 2 (unseen expressions). The baseline uses a frozen encoder with a linear classifier.

6.7. Evaluation Metrics

We report macro-F1 and per-label F1 ($F1_1$, $F1_0$). Thresholds are tuned on development data, and final results are averaged across runs. We additionally compute domain- and register-level error rates for error analysis.

7. Results

7.1. RQ1. How reliably do the models classify euphemisms?

Table 4 presents macro-F1 scores for baseline and fine-tuned models across Split 1 (seen expressions) and Split 2 (unseen expressions).

All models achieve their highest performance in Split 1, where the training, validation, and test sets share euphemistic expressions. In this setting, fine-tuning substantially improves performance for all models. XLM-R and Tohoku BERT achieve the highest macro-F1 (0.73), with large gains over their baselines (from 0.48 to 0.73 and from 0.50 to 0.73, respectively), while Rinna shows a smaller improvement (from 0.45 to 0.68). These results suggest that models benefit from lexical overlap and can learn expression specific patterns when the same euphemisms appear in both training and test data.

Performance drops in Split 2, where all test expressions are unseen. Macro-F1 falls to 0.57–0.61 across models, indicating that generalization to new euphemistic expressions remains challenging. Although fine-tuning still improves performance (e.g., XLM-R from 0.52 to 0.61, Rinna from 0.46 to 0.57), the gains are smaller than in Split 1. This gap suggests that a substantial portion of model performance is driven by lexical familiarity, and that models struggle to interpret euphemistic meaning when surface forms differ.

There are also differences across models. XLM-R shows the largest improvement in Split 1 (+0.25), followed by Tohoku BERT (+0.23) and Rinna (+0.23). In Split 2, XLM-R achieves the highest macro-F1 (0.61), indicating relatively stronger generalization to unseen expressions. Tohoku performs comparably (0.59), while Rinna shows lower performance (0.57), suggesting weaker generalization to new euphemistic expressions.

We observe an apparent per-label performance gap when evaluating on the full canonical test pool, where F1 is consistently higher for euphemistic instances (label 1; F1 ≈ 0.83–0.89) than for literal instances (label 0; F1 ≈ 0.29–0.57) (see Appendix A.3). This gap is partly influenced by test-set composition, as the canonical pool includes always-euphemistic PETs that are labeled only as euphemistic.

To examine this bias, we additionally evaluate on sometimes-euphemistic PETs only, which are label-balanced within each split. Under lexical overlap (Split 1), models achieve relatively balanced performance across labels (e.g., $F1_1$ ≈ 0.75–0.84, $F1_0$ ≈ 0.62–0.80). However, under no-overlap generalization (Split 2), F1 for label 0 drops substantially (≈0.40–0.56), while F1 for label 1 remains higher (≈0.63–0.68). This persistent gap suggests that models tend to over-predict euphemistic usage and rely on lexical associations.

On always-euphemistic PETs, models achieve high F1 for label 1 (≈0.88–0.95), indicating strong performance on lexically stable euphemisms.

Overall, the results suggest that current models capture useful cues for euphemistic usage but rely heavily on lexical familiarity and tend to over-predict positive cases. This highlights the importance of evaluating per-label behavior when assessing pragmatic generalization.

7.2. RQ2. Where do the models produce errors, and why?

To examine model behavior beyond aggregate performance, we conducted error analysis by semantic domain and register. For each domain and register category, we computed the error rate as the proportion of incorrectly classified instances among all test instances in that category:

$$\text{Error Rate} = \frac{\#\text{Incorrect}}{\#\text{Total Instances}}$$

Error rates were calculated separately for each model and split. Because these values represent proportions of misclassified instances, higher values indicate poorer performance within a given category.

Domain	Rinna			Tohoku			XLM-R		
	Base	S1	S2	Base	S1	S2	Base	S1	S2
Aging (46)	57%	13%	27%	38%	13%	31%	54%	8%	19%
Appearance (51)	42%	25%	18%	44%	37%	31%	70%	39%	47%
Bodily Functions (134)	47%	23%	31%	49%	20%	31%	39%	20%	30%
Crime & Social Issues (47)	47%	21%	38%	46%	25%	23%	28%	19%	27%
Death & Dying (208)	51%	10%	16%	35%	7%	18%	40%	9%	16%
Illness & Disability (29)	44%	29%	**39%**	55%	19%	**55%**	29%	19%	35%
Others (28)	40%	11%	14%	53%	7%	11%	60%	14%	7%
Personality (48)	59%	19%	19%	54%	13%	27%	69%	13%	19%
Politics & War (30)	**63%**	**50%**	23%	**78%**	**63%**	50%	**95%**	**67%**	**53%**
Sexuality & Relationships (113)	51%	15%	20%	32%	12%	19%	31%	8%	20%
Workplace & Economy (43)	61%	24%	33%	63%	22%	38%	63%	24%	38%

Table 5: Domain-level **Error Rates** (%) for baseline (Base), Split 1 (S1: seen expressions), and Split 2 (S2: unseen expressions). Lower values indicate better performance and boldface indicates the highest error rate within each column (i.e., vertical comparison).

7.2.1. Per Domain

Table 5 reports domain-level error rates for Splits 1–2. We analyze each split separately to identify which semantic domains are most challenging.

Across models and splits, *Death & Dying* consistently shows the lowest error rates (e.g., 7–10% in Split 1 and around 16–18% in Split 2), indicating that this domain is the most reliably classified. This suggests that euphemisms related to death are highly conventionalized and semantically transparent, making them easier for models to recognize. This domain also contains the largest number of test instances (n=208), which contributes to the stability of the observed error rates.

In contrast, *Politics & War* exhibits the highest error rates across models (e.g., up to 67% in Split 1 and above 50% in Split 2). Errors remain elevated in both splits, indicating that this domain is consistently difficult regardless of lexical overlap. This pattern is largely driven by expressions such as 遺憾 (*ikan*) 'regrettable'. Across models, 遺憾 shows a substantially higher error rate than other items in this domain and accounts for a large proportion of domain-level errors. Removing this expression leads to a noticeable reduction in overall error, suggesting that a small number of politically conventionalized euphemisms disproportionately affect model performance.

The *Illness & Disability* domain also remains challenging, particularly under unseen conditions. Error rates increase substantially in Split 2 for all models (e.g., from around 19% to over 50% for

Register	Rinna			Tohoku			XLM-R		
	Base	S1	S2	Base	S1	S2	Base	S1	S2
Formal (89)	**67%**	17%	17%	**51%**	**24%**	**29%**	49%	**26%**	22%
Informal (82)	27%	13%	18%	36%	10%	18%	28%	12%	24%
Neutral (623)	51%	**19%**	**25%**	45%	17%	27%	**50%**	17%	**26%**

Table 6: Register-level error Rates (%) for baseline (Base), Split 1 (S1: seen expressions), and Split 2 (S2: unseen expressions). Lower values indicate better performance and boldface indicates the highest error rate within each column (i.e., vertical comparison).

Tohoku). Many errors involve orthographic variants such as 障がい (shōgai) 'disability', where partial replacement of kanji with hiragana functions as a mitigation strategy. Instances of 障がい show higher error rates than other items in this domain. Excluding this variant reduces the overall domain error, suggesting that orthographic mitigation poses a distinct challenge for the models.

The *Appearance* and *Workplace & Economy* domains show moderate-to-high error rates, especially in Split 2 (often around 30–40% or higher). In the *Appearance* domain, XLM-R exhibits higher error rates than the monolingual Japanese models, suggesting that culturally grounded or metaphorical expressions may benefit from language-specific pretraining. In both domains, errors are frequently associated with culturally grounded or context-dependent expressions. For example, 社会の窓 (*shakai no mado*; literal meaning: window of society) 'open fly' relies on metaphorical mapping, while 出向 (*shukkō*; literal meaning: go outward) 'forced transfer' can be either literal or euphemistic depending on context. Such expressions show elevated error rates relative to other items and contribute disproportionately to domain-level errors, indicating that metaphor and context dependence remain difficult for the models. Notably, in the *Appearance* domain, the multilingual XLM-R model shows higher error rates than the monolingual models, suggesting that culturally grounded expressions may benefit from language-specific representations.

Overall, the results suggest that models perform well on conventionalized euphemisms but struggle with politically conventionalized, orthographically mitigated, and metaphorical expressions, particularly when lexical cues are absent.

7.2.2. Per Register

We analyze performance across register categories (Formal, Neutral, Informal); see Section 4.3 for details on register annotation. Table 6 reports error rates for baseline and fine-tuned models across Split 1 (seen) and Split 2 (unseen).

A consistent pattern emerges across models: the *Formal* register is the most challenging. Baseline error rates are high (e.g., 67% for Rinna, 51% for Tohoku, and 49% for XLM-R), and although fine-tuning reduces error substantially (e.g., 17% for Rinna in Split 1), performance remains relatively worse than in other registers, especially under unseen conditions (e.g., 29% for Tohoku in Split 2). This suggests that models struggle to capture euphemistic meaning in formal contexts, where indirectness and honorific forms encode pragmatic meaning beyond surface cues.

In contrast, the *Informal* register shows the lowest error rates across models. Fine-tuned models achieve relatively low error (e.g., 12% for XLM-R in Split 1), and baseline performance is also lower than in other registers. However, error increases in Split 2, indicating limited generalization to unseen expressions.

The *Neutral* register shows intermediate but more variable performance. While error is moderate in Split 1, it increases in Split 2 (e.g., up to 27% for Tohoku), suggesting that neutral expressions rely on both lexical and contextual cues and lack consistent surface patterns.

Model differences further highlight sensitivity to register. Tohoku shows higher error in *Formal*, particularly in Split 2, while Rinna remains relatively stable but slightly weaker in *Neutral*. XLM-R exhibits more balanced but less specialized performance across registers.

Overall, these results indicate that euphemism detection depends not only on lexical cues but also on register-specific pragmatic information. Formal expressions remain difficult even after fine-tuning, whereas informal expressions are more easily captured due to their lexical transparency.

7.2.3. Additional Experiment: Tokenization and Pretraining Effects

The higher error rates of Tohoku in the Formal register suggest that performance differences may be partly driven by tokenization and pretraining. While Rinna and XLM-R use SentencePiece, Tohoku relies on MeCab/UniDic-based morphological analysis followed by WordPiece segmentation.

To examine this, we analyzed tokenization fragmentation relative to morpheme boundaries using Sudachi-based segmentation. Formal Japanese expressions often involve *keigo*, which consists of honorific prefixes and auxiliary constructions and thus exhibits complex morphology.

On matched Formal spans shared by all models, Tohoku produced more subword tokens per morpheme than Rinna and XLM-R (paired Wilcoxon tests, $p < 10^{-12}$), with mean differences of +4.5 and +2.4 tokens, indicating greater fragmentation

of morphologically complex expressions. This pattern was particularly evident for keigo forms.

We also observed that Tohoku-specific errors in the matched keigo subset were concentrated in a small number of honorific and euphemistic lexemes (e.g., ご逝去 'passing away', ご年配 'advanced age'). Tohoku further showed higher error rates on keigo expressions compared to the other models.

These findings suggest that Tohoku's elevated error in the Formal register is partly related to tokenization. More fine-grained segmentation may fragment semantically coherent honorific units, making them harder to interpret. However, given the limited data, these results should be interpreted as suggestive rather than conclusive.

8. Discussion

The results provide several insights into how current models interpret Japanese euphemisms.

Model performance is strongly influenced by lexical familiarity. When the same expressions appear in both the training and test sets, all models perform well; however, performance drops substantially for unseen euphemistic forms. This pattern suggests that models rely on memorized lexical patterns rather than inferring the underlying pragmatic functions of euphemisms. The limitation is particularly evident in context-dependent cases, where the distinction between literal and euphemistic meaning depends on cues beyond the sentence. In domains such as *Workplace and Economy*, models frequently misclassify literal descriptions as euphemistic or fail to detect softened references, indicating difficulty in leveraging broader contextual information.

Models also struggle with euphemisms that overlap with formality, politeness, or metaphor. Political euphemisms, for instance, are highly conventionalized and tied to specific genres, leading models to treat them as fixed polite formulas rather than softened criticism, as in 遺憾 (ikan) 'regrettable'. A similar pattern appears with metaphor-based expressions. In the *Appearance* domain, expressions such as 社会の窓 (shakai no mado) 'open fly' rely on visual metaphor to mitigate embarrassment, which requires interpretation beyond the literal meaning. These results suggest that metaphor and pragmatic softening remain challenging for current models.

Orthographic variation introduces an additional challenge. Expressions such as 障がい (shōgai) 'disability', which avoid standard kanji, are often used to convey a gentler impression. However, the models do not appear to treat such variation as a meaningful cue.

Further analysis suggests that tokenization may contribute to model differences, particularly for Tohoku BERT. Honorific expressions (keigo) in Japanese are morphologically complex and encode pragmatic meaning through prefixes and auxiliary constructions. Using Sudachi-based segmentation (Takaoka et al., 2018), we observed that Tohoku produces higher token-to-morpheme ratios on matched formal-register spans, indicating greater fragmentation of these expressions. Errors unique to Tohoku are also concentrated in keigo-related lexemes. While this suggests that segmentation may affect the representation of pragmatically marked forms, the current dataset is insufficient to establish a direct causal relationship.

Finally, XLM-R achieves the strongest performance, particularly in the unseen setting, suggesting that multilingual pretraining may provide broader semantic coverage or exposure to paraphrastic variation. However, in domains involving culturally grounded expressions, monolingual models sometimes perform comparably or better, indicating that language-specific representations remain important for capturing pragmatic nuance.

9. Conclusions

This study examined how transformer-based models interpret Japanese euphemisms under controlled data-splitting conditions designed to isolate lexical familiarity, generalization, and pragmatic inference. The analysis provides a more precise picture of where current systems succeed and where they fail. The results demonstrate that model performance is systematically shaped by domain, register, and pragmatic complexity.

In particular, supplementary analyses suggest that differences in tokenization and morphological segmentation may contribute to variation in performance on formal expressions, especially those involving honorific and euphemistic forms. While these effects remain tentative, they highlight the importance of considering linguistic representation strategies alongside model architecture and training data.

Future work should investigate why multilingual models generalize better than Japanese-only models. This may involve comparing pretraining corpora, analyzing tokenization behavior, or evaluating how deeply each model captures contextual meaning. Examining attention patterns may also help clarify how models process orthographic variation and formulaic expressions.

Another important direction is to extend the dataset to euphemisms that overlap with idioms (e.g., potentially idiomatic expressions (PIEs)), politeness formulas, and honorific forms, allowing systematic comparison between pragmatic and figurative non-literal meaning. Many errors appear

to arise when euphemisms interact with formal or conventionalized linguistic patterns, so incorporating these pragmatic categories would allow a more fine-grained analysis of how different cues influence model behavior.

Finally, adding longer context or document-level data would allow models to distinguish literal and euphemistic readings in ambiguous cases. This is especially important for domains such as Workplace and Economy, where sentence-level context is often insufficient to determine whether an expression softens an unwelcome event.

10. Limitations

A key limitation of this study is the relatively small dataset size and uneven distribution across domains and PET types. While this reflects the natural distribution of euphemisms in corpora, it restricts the ability to draw strong conclusions about domain- or register-specific effects. Future work with larger and more balanced datasets will be necessary to validate these patterns.

Acknowledgements

We would like to thank the Montclair State University NLP Lab for providing the opportunity to conduct this research. This work builds on the lab' s prior studies on euphemism analysis. We also thank the NLP Lab members for their helpful feedback and discussions.

Thank you Julia Sammartino and Patrick Lee for laying the baselines for much of our lab's euphemism work!

11. Bibliographical References

Keith Allan and Kate Burridge. 1991. *Euphemism and dysphemism: Language used as shield and weapon*. Oxford University Press.

Keith Allan and Kate Burridge. 2006. *Forbidden words: Taboo and the censoring of language*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*, 2 edition. Cambridge University Press, Cambridge.

Hasan C. Biyik, Patrick Lee, and Anna Feldman. 2024. Turkish delights: A dataset on turkish euphemisms. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 71–80, Bangkok, Thailand and Online. Association for Computational Linguistics.

Miguel Casas Gómez. 2009. Towards a new approach to the linguistic definition of euphemism. *Language Sciences*, 31:725–739.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Haruko Minegishi Cook. 2006. Japanese politeness as an interactional achievement: Academic consultation sessions in japanese universities. *Multilingua*, 25(3):269–291.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Shoko Hamano. 1998. *The Sound-Symbolic System of Japanese*. CSLI Publications, Stanford, CA.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Patrick Lee, A. Chirino Trujillo, D. Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. Meds for pets: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881. Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, pages 184–190. Association for Computational Linguistics.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic

terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of NAACL-HLT 2022*, pages 4437–4452. Association for Computational Linguistics.

Senko K. Maynard. 1997. *Japanese communication: Language and thought in context*. University of Hawai'i Press.

Steven Pinker. 2003. *The blank slate: The modern denial of human nature*. Penguin Books.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Viking, New York.

Hussein Rababah. 2014. The translatability and use of x-phemism expressions in medical discourse. *Studies in Literature and Language*, 9:1–12.

rinna Co., Ltd. 2023. rinna japanese roberta model. https://huggingface.co/rinna/japanese-roberta-base.

Kazuma Takaoka, Hiroki Ouchi, Toshinori Sakai, Satoshi Akiyama, and Takashi Yamada. 2018. Sudachi: A japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tohoku University. 2020. cl-tohoku/bert-base-japanese-v2. https://github.com/cl-tohoku/bert-japanese.

James Valentine. 1998. Naming the other: Power, politeness and the inflation of euphemisms. *Sociological Research Online*, 3(4):37–53.

A. Appendix

A.1. Per-PET statistics

Statistic	Mean (M)	SD	Min	Max
Sentences per PET	16.7	11.3	3	68
Euphemistic instances	10.8	4.8	1	39
Non-euphemistic instances	5.9	8.8	0	45
PET ambiguity entropy (bits)	0.45	0.48	0.0	1.0
Tokens per sentence†	58.7	13.5	–	–
Lexical density†	0.79	0.06	–	–

Note: Ambiguity was computed as $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$.
† Approximate corpus-level averages.

A.2. Test-set distribution

Domain	PET Type	Label 0	Label 1
Aging	always	0	40
	sometimes	4	4
Appearance	always	0	40
	sometimes	7	4
Bodily Functions	always	0	87
	sometimes	23	24
Crime & Social Issues	always	0	14
	sometimes	18	15
Death & Dying	always	0	158
	sometimes	28	22
Illness & Disability	always	—	—
	sometimes	14	15
Others	always	0	21
	sometimes	3	4
Personality	always	0	36
	sometimes	5	7
Politics & War	always	0	30
	sometimes	—	—
Sexuality & Relationships	always	0	81
	sometimes	17	15
Workplace & Economy	always	0	31
	sometimes	7	5

Note: Counts for sometimes-euphemistic PETs are approximately label-balanced; minor deviations reflect PETs with uneven label distributions.

A.3. Per-label F1 by PET type

Model	Split 1 (lexical overlap)			Split 2 (no overlap)		
	Sometimes		Always	Sometimes		Always
	$F1_1$	$F1_0$	$F1_1$	$F1_1$	$F1_0$	$F1_1$
Rinna RoBERTa	0.75	0.62	0.91	0.64	0.40	0.93
Tohoku BERT	0.82	0.78	0.90	0.63	0.50	0.88
XLM-RoBERTa	0.84	0.80	0.89	0.68	0.56	0.88

Note. Always-euphemistic PETs contain only label 1 instances by definition; therefore, $F1_0$ is not applicable for the always-only subset.