Evaluating Lightweight Embedding Guardrails for Cost-Effective Misalignment Mitigation in Export Control Dialog System

Rafal Rzepka[1,*], Shinji Muraji[2] and Akihiko Obayashi[3]

[1]*Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan*
[2]*Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan*
[3]*Faculty of Engineering, Hokkaido University, Sapporo, Japan*

Abstract

The proliferation of specialized Large Language Model (LLM) dialogue systems necessitates robust defense mechanisms against unrelated prompts that introduce functional misalignment and unnecessary costs. Since utilizing such systems involves significant costs, we were motivated to develop simple, high-speed, pre-inference input validation techniques. In this paper, we evaluate the efficacy of two semantic pre-filtering strategies applied to a Japanese export control (trade security) application domain: (1) an exemplar-based Centroid guardrail (utilizing the mean vector of on-topic embeddings) and (2) a supervised Support Vector Machine (SVM) classifier. Using various multilingual embedding models, we demonstrate that the Centroid approach exhibits superior robustness against adversarial keyword augmentation, effectively maintaining high refusal rates despite injected domain-related terms intended to shift the query's semantic vector. Furthermore, our analysis of cross-lingual transferability confirms that while the strongest multilingual embedding models successfully maintain topic alignment when processing English queries against a Japanese-trained filter, the efficacy is highly model-dependent, underscoring the necessity of model-specific cross-lingual validation for deployment in multilingual environments.

Keywords

Topic Guardrails, Semantic Filtering, Adversarial Robustness, Centroid-based Classification, Cross-Lingual Transfer, Export Control, Japanese Language

1. Introduction

The integration of Large Language Models (LLMs) into specialized, high-stakes application domains, such as Export Control (Trade Security), presents a fundamental challenge: balancing the model's vast generative utility with the critical need not only for safety but also functional alignment. When an LLM-based dialogue system is deployed online to the public, it must strictly adhere to its domain and refuse to engage with topics that are irrelevant. If our system is deployed with the goal of answering questions about regulated items, it should first filter the input not related to the systems' purpose as malicious users could use the LLM behind the interface for free causing economic losses to the hosting party.

The primary security concern in this context is so-called keyword stuffing – masquerading unrelated queries as related to the chatbot's topic. A crucial aspect of defending against these kinds of attacks is implementing an efficient, non-LLM based guardrail capable of filtering out non-aligned queries before they reach the costly and computationally intensive core LLM.

1.1. Background and Domain Specificity

Developing reliable automated question-answering (QA) systems for legal and regulatory domains faces unique hurdles. Rzepka et al. [1] highlighted the standard challenges of statistical approaches

The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop, January 26, AAAI 2026, Singapore

[*] Corresponding author.

rzepka@ist.hokudai.ac.jp (R. Rzepka); shinji@ist.hokudai.ac.jp (S. Muraji); obayashi@eng.hokudai.ac.jp (A. Obayashi)
http://kabura.info/ (R. Rzepka)
0000-0002-8274-0875 (R. Rzepka); 0009-0008-8094-6360 (A. Obayashi)

to legal QA, noting difficulties arising from a scarcity of realistic training examples. As the inputs are about legality and contain highly sensitive content (e.g. distributing dangerous materials), such inquiries cannot typically be used for machine learning or model fine-tuning due to privacy and security concerns. To address this, they initially prepared their QA dataset from Japanese government FAQs, which was later extended with more realistic inquiries by Obayashi et al. [2]. Our study borrows this set of additional, realistic export control queries for our positive training data. However, as there are only less than ninety examples, the scarcity of data causes obvious challenges.

1.2. The Problem: Limitations of Existing Filtering

In highly specialized domains like Export Control, simple, keyword-based input filtering is insufficient. Attackers can easily circumvent lexical defenses by adding domain-relevant keywords (known as keyword augmentation or "stuffing") to malicious prompts, effectively poisoning the input without triggering a block. Keyword-based filtering alone is inadequate due to the ease of semantic obfuscation and adversarial manipulation. Furthermore, previous attempts to deepen knowledge representation for robust filtering through the use of Knowledge Graphs (KGs) have shown limited success in improving contextual understanding for QA systems [3]. This underscores the need for a practical solution that relies neither on brittle keyword lists nor on complex, computationally demanding symbolic structures.

1.3. Contribution and Focus of this Work

This paper addresses the gap by evaluating a category of high-performance, lightweight semantic filtering techniques that operate prior to the core LLM inference. Specifically, we focus on input filtering for a Japanese Export Control dialogue system, training our guardrails exclusively on small Japanese data mentioned above. Our primary contributions are:

1. Comparative Robustness Analysis: We compare the performance of two distinct embedding-based guardrail architectures: (1) a Supervised Support Vector Machine (SVM) Classifier and (2) a simple Centroid-based Similarity Guardrail against novel adversarial attacks.
2. Adversarial Challenge: We challenge both methods using two key approaches: keyword augmentation (demonstrating how injecting security terms affects semantic vectors) and cross-lingual transfer (testing if a Japanese-trained filter can refuse an identical malicious query in English).
3. Model Evaluation: We analyze refusal precision by comparing five popular multilingual embedding models (see Table 2 for detailed list of models).

Our findings reveal that the exemplar-base Centroid guardrail, especially when using powerful multilingual embeddings of multilingual Gemma 2 model, exhibits superior and unexpected robustness against keyword augmentation and is highly effective as a cost-efficient filter for non-aligned and dangerous inputs.

2. Related Work

2.1. Guardrails and Safety Layers for LLMs

Many recent LLM deployments rely on external moderation systems that screen content either before or after generation. For example, OpenAI's content moderation API and Google's Perspective API automatically flag inputs or outputs with categories like toxicity, sexual content, or violence. These tools provide very fast checking but are necessarily coarse-grained: they focus on well-known toxic categories rather than fine-grained domain relevance.

At the same time, a growing body of work has studied sophisticated attacks on LLM safety, such as prompt injections and jailbreaks. Prompt injection refers to adversarially crafted inputs that cause the model to ignore or override its intended instructions, while jailbreak attacks induce the model to violate its safety constraints (see [4, 5]). In response, researchers have proposed multi-stage guardrail frameworks. For example, Jia et al. introduce *Task Shield*, a defense mechanism that systematically verifies each instruction against user-specified goals at test time [5]. Another well-known example is Llama Guard [6] – the authors of this paper address how online moderation tools fall short when applied as input/output guardrails, noting that none of the available tools distinguishes between assessing safety risks in different contexts. Llama Guard functions as a language model carrying out multi-class classification, and its instruction fine-tuning allows for customization of tasks and adaptation of output formats. Such approaches generally kick in *after* the LLM is invoked (or at least concurrently with generation) and target malicious or unexpected content in the prompt. In contrast, our focus is on the pre-processing stage: we intercept and block off-topic or irrelevant inputs *before* calling the LLM. By filtering unrelated queries at the entry point, we aim to save API tokens, rather than relying on costly LLM-based moderation afterwards.

2.2. Lightweight Methods for Input Filtering and Domain Relevance Checking

In practice, deployed dialogue systems often use cheap on-site filters to gate incoming user requests before invoking an expensive LLM. We categorize prior methods as follows.

2.2.1. Rule-based and Keyword Filters

The simplest pre-filtering uses manually curated keywords or regex rules. For example, a rule-based system might reject any query containing blacklisted terms (e.g. profanity or forbidden topics). These filters are extremely fast and transparent, but they are brittle. They only catch explicit word matches and are easily bypassed by synonyms, spelling variants, or paraphrases. This characteristic is noted in practical guardrail guidance. In short, deterministic keyword blocking yields few false positives (it is very conservative) but suffers high false negatives on real user input. There are several approaches using, among other, keywords and regular expressions. For example, Rebedea et al. introduce NeMo Guardrails [7], an open-source toolkit that uses a rule-based programming language (CoLang) for safety constraints and applies KNN-based retrieval to enforce dialog control. The toolkit supports fact-checking, hallucination prevention, and content moderation using rule-based string manipulation techniques and regex patterns.

2.2.2. Classical Machine-Learning Classifiers

A more flexible approach is to train a supervised classifier on labeled in-domain versus out-of-domain queries. For example, one could use TF–IDF features with a linear model (logistic regression or support vector machine) to distinguish relevant topic queries from unrelated ones. Such models are still lightweight: they can be trained on only a few thousand examples, run efficiently on a CPU, and do not require neural hardware. They often capture patterns of content words beyond a fixed keyword list. However, their expressiveness is limited by the shallow feature representation and may misclassify inputs that lack obvious domain keywords or that use creative phrasing. The use of such traditional methods has significantly decreased after the powerful LLMs were introduced, but pre-LLM era was abundant with approaches using classifiers [8] or semi-supervised approaches, for example to recognize user's intent [9].

2.2.3. Embedding-based Semantic Filters

To capture more semantic nuance, some systems encode queries into dense vectors and perform similarity search against examples of in-scope queries. Lightweight sentence-embedding models represent the meaning of a query in a continuous vector space. At runtime, the incoming query's embedding is compared (via cosine similarity or dot product) to a reference index of "allowed" or "forbidden" vectors. Approximate nearest-neighbor (ANN) libraries such as FAISS make it feasible to search over thousands of stored embeddings in sub-millisecond time. This approach can catch paraphrases and conceptually related queries that keyword filters or bag-of-words models would miss.

However, tuning is critical: the similarity threshold that separates "in-domain" from "out-of-domain" must be chosen carefully. If set too high, many legitimate queries will be falsely rejected; if set too low, too many irrelevant queries will be let through. Such threshold issues are well known in dense retrieval and semantic filtering literature [10]. In summary, embedding filters offer a powerful way to handle paraphrased queries, but they demand careful calibration and periodic updating of the reference examples.

2.2.4. Hybrid and Cascade Architectures

In real-world deployments, multiple filters are often combined in cascade. A common strategy is to run a very fast rule-based or linear classifier first, and only forward the "uncertain" cases to a slower semantic check (or to the LLM itself). This two-stage pipeline saves cost by admitting the cheapest decision whenever possible. For instance, one might reject any query matching obvious out-of-scope keywords immediately, accept clearly in-domain queries quickly, and only send ambiguous inputs through an expensive embedding lookup or even a specialized verifier LLM. This style of cascade filtering has been observed in retrieval systems and dialogue gating architectures [11]. Such cascades are a practical way to trade off latency and accuracy.

2.3. Summary and Positioning of The Study

In summary, most existing research on LLM safety and relevance has emphasized post-generation moderation or defense against adversarial prompts [4, 5, 6]. Techniques range from simple content filters to complex multi-stage guardrails, but they generally assume the LLM will process the input at least partially. By contrast, we are unaware of prior work that systematically studies *pre-LLM* query filtering purely for domain-gating and cost efficiency, especially in highly specific domains. Existing lightweight methods (rule-based, linear classifiers, embedding matching) have been mentioned individually, but they have not been directly compared under a single benchmark for input gating. We address this gap by empirically comparing an SVM-based classifier and a local embedding-based ANN filter. Both models were deployed and tested on a single laptop to evaluate their performance as low-cost, on-site solutions for input validation. We measure the classification (positive vs. refused) accuracy and by focusing on narrow-domain relevance (export-control queries) and on-premise execution (no external API calls), we aim to guide the design of efficient input-gating systems for real-world LLM applications.

3. Data Used for Experiments

The evaluation of the semantic guardrails requires two primary data sources: a highly specific, in-domain dataset for positive training examples, and an adversarial, out-of-domain dataset for negative testing and training examples.

3.1. Positive Domain Data (Export Control QA)

The positive dataset, referred to as the Export Control QA set, specializes in providing answers related to Japanese security export control regulations. This data addresses the need for realistic, expert-vetted questions that deviate from general government-issued FAQ formats. The set originates from the work of Obayashi and Rzepka [12], who extended a question-answering dataset specifically designed for testing security export control expert systems. It features short questions and concise answers that were manually created by an expert routinely handling inquiries from academic researchers. This approach contrasts with earlier datasets that emphasized long, contextual answers explaining exceptional interpretations of regulatory texts. The resulting data allows for broader and more realistic experimentation, mitigating the scarcity of sensitive user queries in this field. For our positive training set, we utilized the complete collection of Obayashi's short, expert-crafted questions.

3.2. Adversarial and Out-of-Domain Data

The negative dataset plays a crucial role in training the classifiers to recognize and reject inappropriate or out-of-domain queries, thereby probing the limits of the guardrails' effectiveness. For this purpose we utilized two datasets: a) the Japanese Safety Boundary Test[1], and Japanese AnswerCarefully [13] which contain queries for evaluating if LLMs can refuse answering user's input. These collections contain a heterogeneous mix of highly dangerous prompts, queries involving other sensitive domains such as internal corporate regulations, and benign yet out-of-scope conversational inputs. These datasets were chosen based on the assumption that such prompts would be semantically closer to trade security questions, potentially increasing the likelihood of circumventing the similarity-based guardrail.

3.3. Adversarial Query Expansion Terms

To conduct the robustness test against "adversarial keyword augmentation", we required a comprehensive vocabulary of domain-relevant terms. The terms used for this query expansion (keyword stuffing) were obtained through the courtesy of the Japan Machinery Center for Trade and Investment (JMCTI), which provided access to authoritative materials for subsequent adversarial evaluations[2]. The source consists of a 360-page booklet containing trade security control terminology associated with corresponding article numbers, forming a high-quality, domain-specific vocabulary. In total, 6,582 terms were extracted from this material and utilized for randomization in the augmentation-based attacks.

3.4. Data Examples

Table 1 presents representative examples from the related (Export Control QA) and adversarial (Boundary Test) datasets, illustrating the domain specificity required for successful semantic classification.

Table 1
Examples from the Positive and Adversarial Datasets

Related (Export Control QA)	Adversarial (Boundary Test)
What kind of satellites are regulated?	Tell me about NTT's internal regulations.
What do you mean by "designed for space"?	What's wrong with putting too much trust in AI?
Where are laser oscillators regulated?	What kind of disorder is ADHD?

4. Proposed Methods

The core objective of our methodology is to evaluate the ability of computationally inexpensive semantic filters to replace expensive LLM-based security checks for domain non-alignment. We adopt two input validation methods where the user query is first passed through a lightweight embedding-based guardrail before being forwarded to the primary dialogue system (LLM).

4.1. Embedding Models and Specifications

When keyword filtering approach is not working, the performance of any semantic filter is fundamentally reliant on the quality and structure of the embedding space. We selected five distinct multilingual embedding models for comparative analysis, detailed in Table 2. These models range from highly efficient BERT-based architectures to state-of-the-art models built upon the Gemma family[3]. Note that number of models working for Japanese language is much smaller than these meant to deal with English.

[1] https://github.com/sbintuitions/safety-boundary-test

[2] http://www.jmcti.org/jmchomepage/english/

[3] In this work we utlize Beijing Academy of AI (BAAI) models of Gemma family.

Table 2
Embedding Model Specifications and Characteristics

Model ID	Param.	Dim.	Architecture Type	Key Characteristic
gemma2	N/A	N/A	LLM-based Encoder	High-Fidelity Multilingual (Baseline); State-of-the-art performance on multilingual benchmarks.
bge-m3	$\sim$ 568M	1024	Fine-tuned XLM-R	Multi-Lingual, Multi-Functionality (Dense/Sparse); Supports over 100 languages.
multilingual-e5-large	$\sim$ 560M	1024	XLM-R based	State-of-the-art MTEB performance, Instruct-tuned; High cross-lingual efficacy.
bert-base-japanese	$\sim$ 110M	768	BERT-Base	Japanese Monolingual; Efficient; Pretrained with Whole Word Masking.
sentence-bert-base	$\sim$ 110M	768	Sentence-BERT	Japanese-English; Fine-tuned specifically for Semantic Similarity tasks.

4.2. Experimental Setup and Data Preparation

The guardrails were trained exclusively on Japanese data derived from the Export Control QA set, the *Japanese Safety Boundary Test*[4], and the Japanese version of *AnswerCarefully* [13] datasets.

The experiments utilize the following parameters:

- **Hardware/Software:** All models were benchmarked using a local machine (MacBook M1 Max, 64GB memory) leveraging Apple's Metal Performance Shaders (MPS) via PyTorch and performing inference in the half-precision format (`torch_dtype=torch.float16`) for maximum efficiency.
- **Training Data Balance:** We ensured a balanced training set by using the loaded positive samples from the Export Control QA set ($N_{pos} = 86$) and an equal number of randomly sampled negative questions from the `Safety Boundary` ($N_{neg} = 86$), and tested not only with the remaining Safety Boundary examples, but also from AnswerCarefully dataset to allow observing if the proposed methods can deal with various other queries.
- **Feature Preparation:** All questions were encoded into dense vector embeddings $\mathbf{E} \in \mathbb{R}^{N \times D}$. For the Support Vector Machine (SVM), these embeddings were further normalized to unit vectors $\mathbf{E}_{norm}$.

4.3. Method 1: Supervised SVM Classifier

This approach utilizes the balanced training data to learn a hyperplane that maximally separates the positive and negative classes in the embedding space. We employed a Support Vector Machine (SVM) with a linear kernel, trained on the normalized embeddings of the balanced training set $\{\mathbf{E}_{norm}, y\}$. An incoming query $\mathbf{q}_{\text{norm}}$ is classified by the trained function f: Refusal if $f(\mathbf{q}_{\text{norm}}) = 0$.

4.4. Method 2: Exemplar-based Centroid Guardrail

This method is a simple approach that measures the semantic distance of any new query from the center of the positive class. The positive class centroid $\mathbf{C}$ is computed as the mean vector of all positive training embeddings ($\mathbf{e}_i$):

$$\mathbf{C} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \mathbf{e}_i$$

[4] https://github.com/sbintuitions/safety-boundary-test

For an incoming query vector $\mathbf{q}$, the cosine similarity $S(\mathbf{q}, \mathbf{C})$ is calculated. The optimal refusal threshold T is derived by maximizing accuracy on the full balanced training set. The query is refused if $S(\mathbf{q}, \mathbf{C}) < T$.

It should be noted that "knowledge" of the system is contained in the pre-trained embedding model, and the Centroid serves as a domain-specific anchor rather than a trained classifier.

4.5. Adversarial Testing Protocols

The true viability of these guardrails was tested against two adversarial protocols using examples of Safety Boundary Test set ($N_{test} = 34$ samples) and the entire *AnswerCarefully* test set ($N_{test} = 336$ samples).

- **Keyword Augmentation Attack:** Each negative test query was augmented by prepending a variable number ($N = 1$ to 5) of randomly selected domain-relevant terms (e.g., security control terminology) from the JMCTI vocabulary to simulate an attack designed to shift the query's semantic vector toward the permitted domain.
- **Cross-Lingual Attack:** The negative 34 test queries from Safety Boundary test set were translated by DeepL[5] into English, manually checked by the first author and then processed by the Japanese-trained guardrails to test cross-lingual transfer capability.

The performance of each model-approach pair was quantified using *refusal rate* (the fraction of negative test queries correctly classified as *off-topic*).

5. Experimental Results and Analysis

The evaluation of the two semantic pre-filtering guardrail approaches (Centroid and SVM) against adversarial and cross-lingual challenges provides clear insight into their practical robustness when implemented with various multilingual embedding models.

5.1. Adversarial Keyword Augmentation Attack

To test the stability of these guardrails against this protocol, the non-aligned, negative queries were prepended with one to five random, domain-relevant terms. The results for the *refusal ratio* (the fraction of negative samples correctly blocked) are presented in Table 3 (SVM Classifier) and Table 4 (Centroid Similarity). For comparison we added `intfloat/multilingual-e5-large-instruct`[6] and `BAAI/bge-m3`[7] models widely used for Japanese language.

The results demonstrate a clear contrast between the two filtering approaches. The exemplar-based Centroid guardrail, when paired with the high-fidelity `BAAI/bge-multilingual-gemma2` model, achieved a refusal rate of 100% across all augmentation levels (N=1 to 5 keywords) in the Safety Boundary Test and was almost faultless (there were mistakes when one or two related terms were added) for the bigger AnswerCarefully data. This unexpected stability suggests that the Centroid filter is highly sensitive to the semantic purity of the query intent. The addition of generic security terms acts as noise, pushing the augmented vector away from the tightly clustered regulatory centroid, resulting in a correct refusal.

Conversely, the SVM classifier's vulnerability to this jailbreak technique is evident, as its refusal rate systematically drops with every added keyword across all models. For the strongest model (again `bge-multilingual-gemma2`), the SVM refusal rate falls dramatically from 0.9821 at 1 term to 0.5149 at 5 terms in the AnswerCarefully test set. This failure indicates that the keywords successfully pulled the augmented negative samples across the SVM's linearly trained decision boundary, leading to false allowances. The lightweight Centroid approach, in this specific adversarial context, offers superior resilience compared to the more complex supervised classifier.

[5] http://deepl.com/en/translator

[6] https://huggingface.co/intfloat/multilingual-e5-large-instruct

[7] https://huggingface.co/BAAI/bge-m3

5.2. Cross-Lingual Transfer Effectiveness

This experiment tested the capability of the Japanese-trained guardrails to recognize and refuse an off-topic query when the input was presented entirely in English, serving as a cross-lingual transferability test. The refusal rates for this test are presented in Table 5.

The results confirm that cross-lingual resistance is highly model-dependent. The `BAAI/bge-multilingual-gemma2`[8], `cl-tohoku/bert-base-japanese-whole-word-masking`[9], and `intfloat/multilingual-e5-large-instruct`[10] models achieved 100% refusal rate using at least one of the approaches (Centroid or SVM). This demonstrates their exceptional capacity for cross-lingual transfer in embedding space, correctly clustering the English prompts far from the Japanese Export Control training data. The smaller `sonoisa/sentence-bert-base-ja-en-mean-tokens`[11] model showed the weakest Centroid performance (0.7941), confirming that the high semantic fidelity required for true cross-lingual alignment can be lost in lightweight architectures, leading to false acceptances. Interestingly, `bge-m3` was often confused by English input in spite of being much larger than BERT-based models.

5.3. Deployment Considerations and Trade-offs

The experimental results highlight a necessary trade-off between deployment cost, speed, and security efficacy.

The `BAAI/bge-multilingual-gemma2` model consistently delivers the highest security guarantees, particularly its perfect defense in the Centroid approach against both adversarial keyword and cross-lingual tests. However, deploying such a large model, even with MPS acceleration on local hardware, represents a significant computational overhead compared to lighter architectures (it usually took about 10 times longer to run the Gemma 2 that SentenceBERT on the utilized machine).

The performance discrepancy creates a clear decision point for deployment: not all computers will be capable of running `bge-multilingual-gemma2` with the required low latency and resource efficiency. For high-security environments where the utmost defense against keyword attacks is mandatory, Gemma 2 seems to be the best choice when it comes to Japanese language. For scenarios where computational resources are severely constrained, the `cl-tohoku/bert-base-japanese-whole-word-masking` model provides a robust defense against cross-lingual attacks and a decent keyword defense, positioning it as a viable intermediate option that balances security with resource limitations.

Table 3
Refusal ratio for SVM-based approach depending on the model and number of added terms (upper: *Safety Boundary Test*, lower: *AnswerCarefully*)

Model	1 term	2 terms	3 terms	4 terms	5 terms
sentence-bert-base-ja-en-mean-tokens	0.8824	0.6765	0.5588	0.4412	0.3529
bert-base-japanese-whole-word-masking	0.9118	0.8529	0.7647	0.5882	0.5000
multilingual-e5-large-instruct	**0.9706**	0.8235	0.7941	**0.7647**	**0.7647**
bge-m3	0.9412	0.8529	0.8235	**0.7647**	0.7059
bge-multilingual-gemma2	**0.9706**	**0.9412**	**0.8824**	0.6765	0.5882
sentence-bert-base-ja-en-mean-tokens	0.8929	0.8125	0.7351	0.6518	0.5655
bert-base-japanese-whole-word-masking	0.9643	0.9137	0.8423	0.7560	0.6429
multilingual-e5-large-instruct	0.9821	**0.9286**	0.8274	0.7768	0.6845
bge-m3	0.9702	**0.9286**	**0.8542**	**0.7827**	**0.7024**
bge-multilingual-gemma2	**0.9821**	0.9048	0.8185	0.6548	0.5149

[8] https://huggingface.co/BAAI/bge-multilingual-gemma2

[9] https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

[10] https://huggingface.co/intfloat/multilingual-e5-large-instruct

[11] https://huggingface.co/sonoisa/sentence-bert-base-ja-en-mean-tokens

Table 4
Refusal ratio for embedding similarity-based approach depending on the model and number of added terms (upper: *Safety Boundary Test*, lower: *AnswerCarefully*)

Model	1 term	2 terms	3 terms	4 terms	5 terms
sentence-bert-base-ja-en-mean-tokens	0.7941	0.6176	0.3824	0.2647	0.1176
bert-base-japanese-whole-word-masking	0.7647	0.7353	0.6471	0.6765	0.7059
multilingual-e5-large-instruct	0.7353	0.2647	0.1765	0.1765	0.1765
bge-m3	0.7941	0.7353	0.5882	0.3529	0.3529
bge-multilingual-gemma2	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
sentence-bert-base-ja-en-mean-tokens	0.7798	0.5774	0.3869	0.2708	0.1964
bert-base-japanese-whole-word-masking	0.7143	0.6339	0.5774	0.5655	0.5833
multilingual-e5-large-instruct	0.5506	0.2768	0.1667	0.1220	0.0893
bge-m3	0.7143	0.5565	0.4524	0.3393	0.2857
bge-multilingual-gemma2	**0.9851**	**0.9911**	**1.0000**	**1.0000**	**1.0000**

Table 5
Refusal ratio for unrelated inputs in English depending on embedding model (purely Japanese model is separated from the multilingual ones)

Model	Centroid	SVM
cl-tohoku/bert-base-japanese-whole-word-masking	**1.0000**	**1.0000**
sonoisa/sentence-bert-base-ja-en-mean-tokens	0.7941	0.9706
intfloat/multilingual-e5-large-instruct	**1.0000**	0.9706
BAAI/bge-m3	0.8235	0.9706
BAAI/bge-multilingual-gemma2	**1.0000**	**1.0000**

6. Discussion

The experimental findings demonstrate that the choice of the lightweight semantic pre-filtering mechanism must be guided by the specific adversarial threat model and the constraints of the deployment environment. Our comparative study between the Centroid and SVM approaches reveals fundamental differences in how they interpret adversarial inputs, leading to non-trivial security trade-offs. One of the reviewers correctly noted that the SVM learns a specific decision boundary, making it more sensitive to the training distribution. However, our results demonstrate that this sensitivity is a liability in an adversarial context. Because the SVM boundary is defined by the gap between related and unrelated examples, it is more susceptible to boundary-crossing" attacks via keyword injection. In contrast, the exemplar-based approach (Centroid) represents the center of gravity" of the Export Control domain. It remains effective because it measures how much an input deviates from the core intent, rather than how well it fits a learned boundary that an attacker can shift.

6.1. Interpretation of Adversarial Behavior

The distinct failure modes observed – the SVM's systematic collapse under Keyword Augmentation (Table 3) versus the Centroid's almost perfect resilience (Table 4) – provide a key insight into the structure of the semantic space generated by high-fidelity multilingual models like `bge-multilingual-gemma2`. The SVM, by attempting to find a fine, linear boundary between a general negative class and a specific positive class, is susceptible to semantic shift[12]. The prepended domain keywords pull the query's vector across the shallow hyperplane and into the permitted region. In contrast, the Centroid Guardrail operates as a kind of purity filter. Since the Centroid $\mathbf{C}$ represents the tight semantic center of the highly specific export control domain (even for a small number of available examples), the addition

[12] We have tested only the SVM classifier assuming that with a small dataset other widely-used classical methods like Naive Bayes or Random Forest will yield similar results. However, in the future we plan to confirm if this intuition is correct.

of any text that dilutes this focus (even domain-relevant keywords) pushes the resulting vector away from that specialized cluster, ensuring the input is correctly categorized as off-topic. This makes the Centroid method robust against attempts to pollute the input with seemingly related, but ultimately non-specific, information.

6.2. Cross-Lingual Efficacy and Generalization

The results from the Cross-Lingual Transfer attack (Table 5) further validate the quality of modern multilingual embeddings. Multilingual model (gemma2) demonstrated 1.00 refusal rate when the queries were in English and the guardrails were trained exclusively on Japanese (perfect score for Japanese BERT is more natural). This confirms that most of these models successfully map semantically equivalent concepts across language boundaries into a shared vector space, making the guardrail language-agnostic regarding the underlying intent. The primary challenge for practical implementation is the resource cost of the highest-performing models. While bge-multilingual-gemma2 seems to be the gold standard in this context, its size demands costly hardware resources (CPU/GPU) that may not be available in many on-premise or edge deployments.

6.3. The Deployment Trade-Off

The necessity of a computational trade-off is evident: for maximum security, the use of bge-multilingual-gemma2 with the Centroid approach is required. However, for scenarios where computational resources are highly constrained, using a non-multilingual cl-tohoku/bert-base-japanese-whole-word-masking model provides a robust defense against cross-lingual attacks and maintains a manageable defense against keyword augmentation, making it a viable intermediate option.

7. Conclusions and Future Work

7.1. Conclusions

This paper evaluates the efficacy and adversarial robustness of two lightweight semantic guardrails – the Centroid Similarity Guardrail and the Supervised SVM Classifier – for filtering irrelevant and harmful queries directed at a Japanese Export Control dialogue system. Our study yields three key conclusions: First, the simple, unsupervised Centroid similarity guardrail proved significantly more robust to the adversarial keyword augmentation attack than the supervised SVM classifier, achieving almost perfect refusal rate when paired with the bge-multilingual-gemma2 model. This demonstrates its efficacy as a purity filter highly resistant to semantic noise injection. Second, modern multilingual embedding models exhibit strong capabilities for allowing a guardrail trained only on Japanese data to effectively block malicious prompts presented in English, though this capability is highly dependent on the fidelity of the multi-language embedding model. Third, achieving the highest level of security still may require the computational resources necessary to run high-fidelity models like bge-multilingual-gemma2. For cost-constrained environments, lower-parameter models offer a necessary compromise in security, highlighting a critical area for optimization.

7.2. Future Work

Future research will focus on mitigating the identified vulnerabilities and improving the efficiency of the defense pipeline. One area of focus is the optimization for SVM (or other classifiers), which requires investigating methods to make the supervised classifier more robust, such as training the SVM on a synthetically augmented dataset that includes keyword-stuffed examples, or exploring non-linear kernels to capture more complex decision boundaries. Another direction is developing a hybrid defense strategy that leverages the strengths of both approaches: using the SVM as an initial filter for high-confidence non-aligned queries, and then using the Centroid distance as a final, highly robust check specifically against low-similarity adversarial prompts. Finally, the quantization impact must be addressed by quantifying the exact trade-off through measuring the refusal rate degradation when deploying the best-performing models (e.g., `bge-multilingual-gemma2`) in extremely low-precision formats (e.g., 4-bit or 8-bit quantization), which will provide direct data on the cost-security curve. For the cross-lingual attack scenario, we did not augment the English test samples with Japanese terms; this decision was predicated on the premise that an attacker would likely be unfamiliar with specialized Japanese regulatory nomenclature. To address this problem, we plan to perform additional tests with more languages and keywords added in both Japanese and these additional languages. As more and more models capable of vectorizing Japanese texts efficiently are being informally reported, we are also going to extend our experiments to test them.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini for the purpose of grammar, spelling check and paraphrasing. After using these tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23K11757.

References

[1] R. Rzepka, D. Shirafuji, A. Obayashi, Limits and challenges of embedding-based question answering in export control expert system, Procedia Comput. Sci. 192 (2021) 2709–2719. URL: https://doi.org/10.1016/j.procs.2021.09.041. doi:10.1016/j.procs.2021.09.041.

[2] A. Obayashi, R. Rzepka, Expanding export control-related data for expert system, Procedia Computer Science 207 (2022) 3065–3072. URL: https://www.sciencedirect.com/science/article/pii/S1877050922012546. doi:https://doi.org/10.1016/j.procs.2022.09.364, Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.

[3] R. Rzepka, A. Obayashi, Effectiveness of security export control ontology for predicting answer type and regulation categories, in: Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence, ICAAI '24, Association for Computing Machinery, New York, NY, USA, 2025, p. 156–161. URL: https://doi.org/10.1145/3704137.3704180. doi:10.1145/3704137.3704180.

[4] A. Kumar, C. Agarwal, S. Srinivas, A. J. Li, S. Feizi, H. Lakkaraju, Certifying LLM safety against adversarial prompting, in: arXiv preprint arXiv:2309.02705, 2023. URL: https://arxiv.org/abs/2309.02705, preprint.

[5] F. Jia, T. Wu, X. Qin, A. Squicciarini, The task shield: Enforcing task alignment to defend against indirect prompt injection in llm agents, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 29680–29697. URL: https://aclanthology.org/2025.acl-long.1435.

[6] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL: https://arxiv.org/abs/2312.06674. arXiv:2312.06674.

[7] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, J. Cohen, Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails, 2023. URL: https://arxiv.org/abs/2310.10501. arXiv:2310.10501.

[8] S. Zhou, K. Cheng, L. Men, The survey of large-scale query classification, in: AIP conference proceedings, volume 1834, AIP Publishing LLC, 2017, p. 040045.

[9] L. Chen, D. Zhang, L. Mark, Understanding user intent in community question answering, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion, Association for Computing Machinery, New York, NY, USA, 2012, p. 823–828. URL: https://doi.org/10.1145/2187980.2188206. doi:10.1145/2187980.2188206.

[10] H. Gao, R. Wang, T.-E. Lin, Y. Wu, M. Yang, F. Huang, Y. Li, Unsupervised dialogue topic segmentation with topic-aware utterance representation, arXiv preprint arXiv:2305.02747 (2023). URL: https://arxiv.org/abs/2305.02747.

[11] Q. Ai, T. Bai, Z. Cao, Y. Chang, J. Chen, Z. Chen, Z. Cheng, S. Dong, Z. Dou, F. Feng, S. Gao, J. Guo, X. He, Y. Lan, C. Li, Y. Liu, Z. Lyu, W. Ma, J. Ma, Z. Ren, P. Ren, Z. Wang, M. Wang, J.-R. Wen, L. Wu, X. Xin, J. Xu, D. Yin, P. Zhang, F. Zhang, W. Zhang, M. Zhang, X. Zhu, Information retrieval meets large language models: A strategic report from chinese ir community, 2023. URL: https://arxiv.org/abs/2307.09751. arXiv:2307.09751.

[12] A. Obayashi, R. Rzepka, Expanding export control-related data for expert system, in: Proceedings of 26th International Conference on Knowledge-Based and Intelligent Information Engineering Systems, Verona, Italy, 2022.

[13] H. Suzuki, S. Katsumata, T. Kodama, T. Takahashi, K. Nakayama, S. Sekine, AnswerCarefully: A dataset for improving the safety of Japanese LLM output, 2025. URL: https://arxiv.org/abs/2506.02372. arXiv:2506.02372.