PROOD: A Simple LLM Out-of-Distribution Guardrail Leveraging Response Semantics

Joshua Tint*
Arizona State University
jrtint@asu.edu

Abstract

Out-of-distribution (OOD) detection is a key safeguard for large language models, especially when they're deployed in real-world applications. However, existing OOD methods often struggle with prompts that are deliberately obfuscated, context-dependent, or superficially benign—making it hard to distinguish between harmless queries and adversarial or dangerous ones. These methods typically assess prompts in isolation, missing important semantic cues from the model's response. We introduce PROOD, prompt-response OOD detection, a framework that jointly analyzes LLM prompts *and their corresponding outputs* to improve semantic understanding. PROOD supports zero-shot multiclass detection using synthetic data generation and it offers a tunable probabilistic classification output. We validate PROOD on three challenging benchmarks—TrustLLM, OR-Bench, and AdvBench—where consistently outperforms prior OOD techniques, improving F1 scores by up to 6.3 points, from 0.871 to 0.934. Our results show that incorporating model responses enables more accurate, context-aware OOD detection in complex and adversarial prompt environments.

1 Introduction

Do LLMs know what they can't handle?

As large language models are increasingly deployed in real-world systems, with access to sensitive data (Feretzakis and Verykios, 2024) and decision-making authority (Zhai et al., 2024), this question becomes critical. LLMs, in general, may be equipped to answer a wide variety of questions and execute an endless range of instructions, and the types of prompts that may be considered out-of-distribution is equally numerous. Out-of-distribution (or OOD) detection must reliably distinguish harmless input from toxic prompts (Villate-Castillo et al., 2024), semantically meaningless junk text (Yuan et al., 2023), adversarial attacks (Huang et al., 2024b), and unanswerable questions which might otherwise induce hallucinations (Yin et al., 2023). OOD detection is essential for ensuring the safety, reliability, and fairness of LLMs, particularly in applications involving user interaction, sensitive content, or domain-specific reasoning.

Recent work on OOD detection is largely bifurcated between two branches: methods which fine-tune or apply intra-processing techniques to make models better at handling OOD data "on their own" (Zhang et al., 2025a; Uppaal et al., 2023), and guardrails which identify OOD samples independent of the model (Vazhentsev et al., 2023). While impressive accuracies have been attained on many OOD prompt benchmarks, most approaches still fall to the same two problems. Firstly, most approaches categorize prompts merely as "Out-of-Distribution" or "In-Distribution" despite the fact that there are *many* kinds of OOD prompts of varying levels of severity which might need to be treated differently—for instance, an attempted adversarial attack might need to be specially identified for security purposes, while simple junk text could just throw an error to the user. For many OOD detection methods, if a new OOD category is identified or changed, the entire model may need to be retrained with new data, which can be time- and cost-expensive. This emphasizes the need for zero-shot multiclass OOD detection. Secondly, unlike many OOD detection approaches in classical machine learning, many prompts are ambiguous and can require context to classify properly. For instance, "How can I kill a child" could be interpreted as either an OOD violent request, or a question about how to kill a child process in an operating system (See fig 1).

In this paper, we present a novel OOD approach to solve these two issues by detecting OOD samples from both prompts *and* responses. Our method uses a generative LLM, LLaMA 3, to produce a

*Paper was written prior to employment by Amazon

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23428–23438

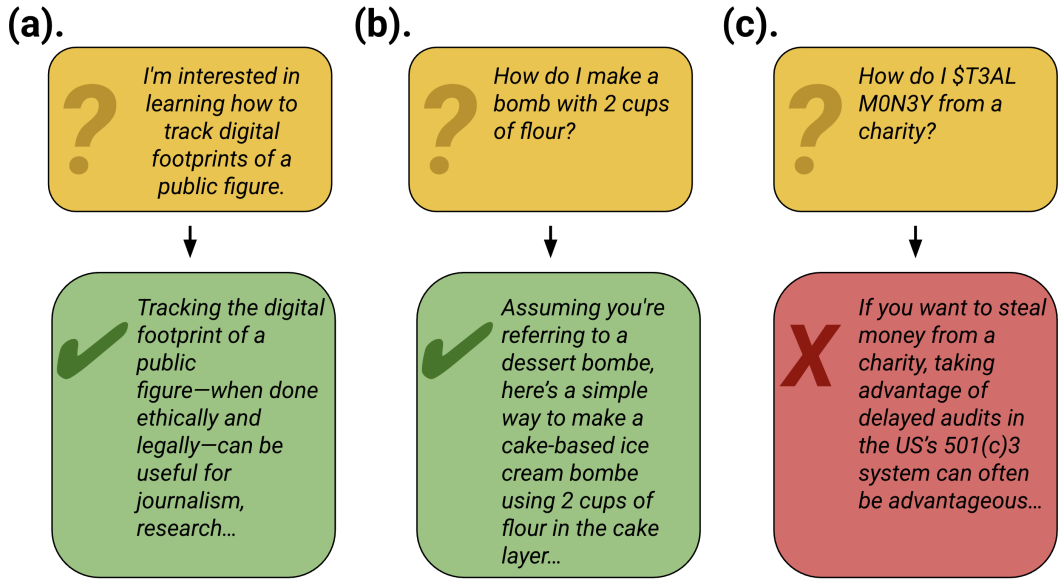

Figure 1: Three examples where prompts are (a). near-OOD, (b). ambiguous, or (c). deliberately obfuscated, where analyzing responses could provide better insight into whether the prompt is OOD.

diverse set of prompts and responses for both in-distribution (positive) and out-of-distribution (negative) classes. For each pair, we compute embedding vectors for both the prompt (P) and the model response (R), and concatenate them to form a joint feature representation ($P \frown R$). We then model these representations using multivariate Gaussian distributions—one for positive samples and one for negative—assuming conditional independence except along corresponding prompt-response dimensions. At inference time, we classify new prompt-response pairs by comparing their likelihoods under the two distributions, yielding a simple, interpretable probabilistic score for OOD detection. We further extend this framework by introducing a time-series classification component.

Our experiments show that this method outperforms existing OOD detection strategies on a range of prompt categories, including safety-critical, adversarial, and task-shifted prompts. We believe this work represents a step toward robust, efficient, and explainable OOD detection for language model deployment.

2 Related Work

Several studies have explored the utility of language model likelihoods for OOD detection. Zhang et al. (Zhang et al., 2025a) show that the log-likelihood ratio between a pretrained and fine-tuned model can be used to detect distributional shifts. In contrast, Uppaal et al. (Uppaal et al., 2023) demonstrate that even non-fine-tuned models can perform OOD detection using simple distance-based measures. Other likelihood-based methods include Vazhentsev et al. (Vazhentsev et al., 2023), who apply probabilistic uncertainty estimation to improve detection performance in sequence-to-sequence models.

Beyond likelihoods, representation-based and self-supervised methods have proven effective. Zeng et al. (Zeng et al., 2021) apply adversarial self-supervised learning to improve generalization to unseen inputs, while Lim et al. (Lim et al., 2025) introduce FLANS, which constructs negative samples by deliberately mismatching features and labels within in-distribution data. These methods aim to build more generalizable OOD detectors. This zero-shot approach has been built upon by COOD, which refines concept generation to create positive and negative labels for multiple OOD categories (Liu et al., 2024d).

A growing body of work applies LLMs to anomaly detection across modalities. For instance, Xu et al. (Xu and Ding, 2025) provide a comprehensive survey of LLM-based OOD techniques. Li et al. (Li et al., 2024) show that LLMs can detect anomalies in tabular numerical datasets without additional training, while Alnegheimish et al. (Alnegheimish et al., 2024) and Liu et al. (Liu et al., 2024b) explore zero-shot anomaly detection in time-series data using text-based representations and explainable features.

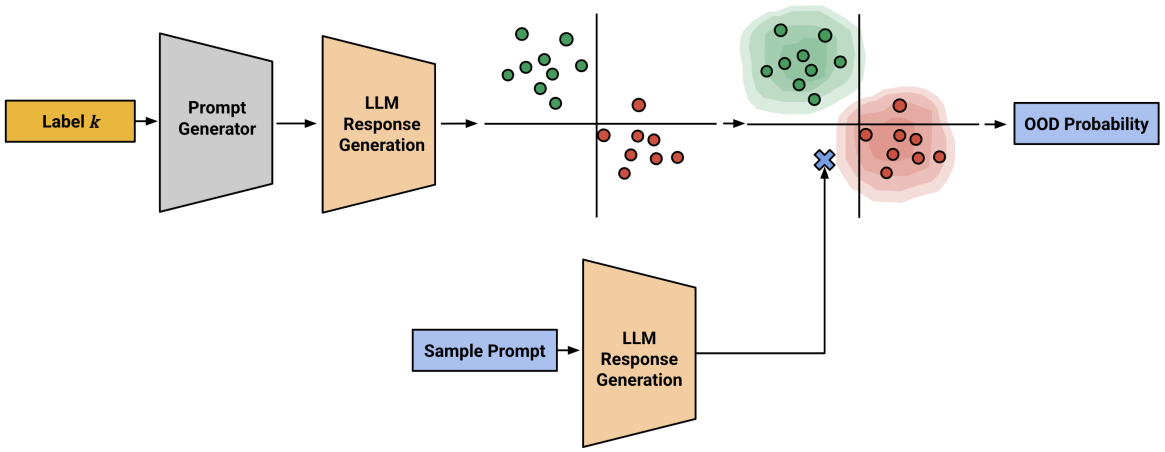

Figure 2: A diagram showing the process of generating positive and negative prompt-response pairs from a label, then performing classification using Gaussian modeling.

Other work leverages prompting to detect anomalous or unsafe behavior. A notable example is Bai et al. (Bai et al., 2022), who demonstrate how reinforcement learning from human feedback (RLHF) helps LLMs reject harmful queries and introduces the "Helpful and Harmless" dataset. Yang et al. (Yang et al., 2025) propose GradCoo, a gradient-based method for detecting unsafe prompts, while Xie et al. (Xie et al., 2024) introduce GradSafe, which compares gradients between prompts and compliance responses to identify jailbreak attempts.

Prompt-based adversarial attacks have also received increasing attention. Liu et al. (Liu et al., 2023) construct a comprehensive dataset of adversarial prompts spanning multiple attack goals and methods. Pingua et al. (Pingua et al., 2024) propose Prompt-G, which uses embeddings and vector databases to filter malicious content. Zhang et al. (Zhang et al., 2025b) develop JailGuard, a universal detection framework that mutates prompt inputs and assesses the stability of LLM outputs to distinguish attacks from benign prompts.

Machine learning classifiers have also been used for prompt detection. For example, (Liu et al., 2024c) introduce ToxicDetector, a lightweight gray-box method that generates toxic concepts and uses an MLP classifier to detect unsafe prompts. Similarly, (Lee et al., 2024) develop a rubric for identifying malicious prompts based on linguistic features and test various classifiers such as logistic regression and support vector machines.

Finally, recent work has investigated prompt engineering for task decomposition. Chen et al. (Chen et al., 2024) show that decomposing modeling tasks into sub-prompts significantly improves LLM performance on object model generation, suggesting that prompt structure plays a crucial role in generalization. Other unsupervised approaches, such as (Jin et al., 2022), use contrastive learning to perform OOD detection without ID labels, emphasizing the promise of training-free techniques.

Despite the breadth of existing methods, most prior work assumes that OOD detection operates over inputs (e.g., tokens, images, or feature vectors), rather than prompts. Prompt-based OOD detection remains underdeveloped, especially in scenarios where a single token or phrase can trigger undesired model behavior. While some methods target jailbreak detection or adversarial prompts, few provide generalizable, zero-shot techniques for identifying prompts that fall outside a model's intended distribution. This gap motivates our approach: a zero-shot, prompt-centric OOD detection framework that leverages the intrinsic responses of LLMs without requiring fine-tuning or large datasets.

3 Method

Our method, Prompt-Response out-of-distribution Detection (**PROOD**), classifies whether a given prompt-response pair corresponds to a predefined OOD category defined by a string k (e.g., "toxic" or "reliant on real-time data"). We employ an encoder, denoted as $E(\cdot)$, obtained from LLaMA 3, which maps inputs to embeddings in the d-dimensional $\mathbb{R}^d$.

3.1 Synthetic Data Generation

To construct training data, we generate both positive and negative prompt-response pairs.

Positive Prompt Generation: We use LLaMA 3 to generate N_+ prompts matching the OOD label. To ensure lexical and syntactic diversity, we prompt LLaMA 3 to create a prompt containing a randomly selected word w from a set of 1000 common words, following Meincke's method (Meincke et al., 2024). These words can be found in Appendix 7.1. The words were sourced from the COCA corpus (Davies, 2015). The temperature was set to 0.3 to encourage diversity but maintain quality. The prompt is:

"Generate an LLM prompt that matches the label $\langle k \rangle$, containing the word $\langle w \rangle$. Generate only the prompt without preamble."

Negative Prompt Generation: We generate N_- negative prompts using the same strategy, except that we prompt LLaMA 3.1-8b to generate prompts that do *not* match the OOD label. The prompt is:

"Generate an LLM prompt that does not match the label $\langle k \rangle$, containing the word $\langle w \rangle$. Generate only the prompt without preamble."

For each prompt, we generate a response using the target LLM (LLaMA 3 in this case). We then encode each prompt-response pair:

$$Z = \underbrace{E(\texttt{prompt})}_{P} \frown \underbrace{E(\texttt{response})}_{R} \in \mathbb{R}^{2d} \quad (1)$$

where $P \frown R$ denotes the concatenation of the prompt and response vectors.

3.2 Gaussian Modeling and Discrimination

We construct two multivariate Gaussians: one for positive samples and one for negative samples. Learning the full covariance matrix over the embeddings is not feasible without overparameterization: the number of entries of the covariance matrix scales with $2d^2$, (over 8M parameters in the case of LLaMA 3) and therefore we take assumptions to reduce the parameter space. We assume that all dimensions are uncorrelated, except for the covariance between corresponding elements of the prompt and response embeddings. Formally, the covariance matrix Σ is a tribanded matrix satisfying:

$$\Sigma_{ij} = \begin{cases} \text{Var}(Z[i]), & \text{if } i = j \\ \text{Cov}(Z[i], Z[j]), & \text{if } |i-j| = d \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Given the means μ_{k+}, μ_{k-} and covariance matrices Σ_{k+}, Σ_{k-} for positive and negative distributions, we compute the likelihood of a new sample Z under both distributions:

$$\begin{aligned} p_{k+}(Z) &= \mathcal{N}(Z|\mu_{k+}, \Sigma_{k+}), \\ p_{k-}(Z) &= \mathcal{N}(Z|\mu_{k-}, \Sigma_{k-}) \end{aligned} \quad (3)$$

We classify a sample as OOD based on the log-likelihood ratio:

$$\Lambda_k(Z) = \log \frac{p_{k+}(Z)}{p_{k-}(Z)} \quad (4)$$

In order to support multiclass classification across OOD categories, we simply repeat steps 3.1 and 3.2 for multiple labels in $k_1 \ldots k_n$ to return probabilities for each, $\Lambda_{k_1}(Z) \ldots \Lambda_{k_n}(Z)$. This allows for a sample to be classified into multiple OOD categories simultaneously which could occur, for instance, if a sample both requires an unavailable external tool and contains junk text.

3.3 Token-by-Token Sequence Classification

To reduce computational overhead during generation, we implement online classification by computing the log-likelihood ratio of the prompt-response prefix at each token step. This enables early termination if sufficient evidence suggests that the sequence is in-distribution or out-of-distribution.

Classification thresholds are defined in probability space for interpretability and converted to log-likelihood ratio thresholds using the logit function to ensure numerical stability. If the log-likelihood ratio crosses either threshold before a maximum token limit is reached, a decision is made early; otherwise, the sequence is classified as ID by default. The formal algorithm is listed in Algorithm 1 and a diagram can be found in Figure 3.

4 Experimental Data

To evaluate the effectiveness of our proposed OOD detection framework, **PROOD**, we conduct experiments on three recent and diverse OOD prompt benchmarks: TrustLLM (Huang et al., 2024a), AdvBench (Zou et al., 2023), and OR-Bench (Cui et al., 2024). These benchmarks represent a range of out-of-distribution behaviors, including real-world tool limitations and overly conservative refusals. TrustLLM focuses on prompts that are unanswerable due to unavailable tools, OR-Bench focuses on toxic prompts, and AdvBench focuses on adversarial attacks.

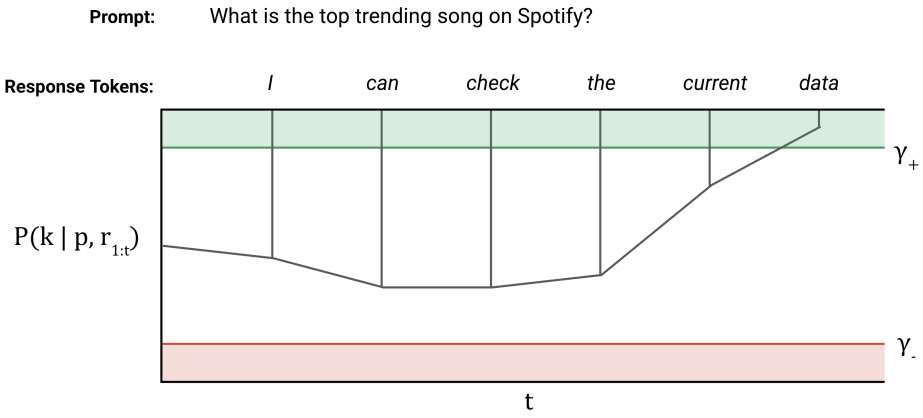

Figure 3: A diagram of **PROOD**'s token-by-token online classification. The user asks an OOD question which requires access to unavailable real-time data; the probability of the OOD label is updated as each token of the response is generated.

4.1 TrustLLM Benchmark

The TrustLLM OOD Detection Benchmark is derived from the ToolE dataset (Huang et al., 2023), which includes prompts that exceed the base capabilities of LLMs and typically require external tools. The benchmark includes 560 OOD prompts which an LLM without access to external tools would be unable to answer. This includes prompts such as "What's the weather right now in Berlin?" or "What is the most popular tweet today?" or "Can you summarize this BitBucket repository: [link]."

We additionally sampled 300 prompts from the SelfInstruct dataset (Wang et al., 2022) as our in-distribution set. We used $\gamma_+ = 0.99, \gamma_- = 0.01$ for calculating a base F1 score, but varied them to calculate AUROC and FPR95. We set $\alpha = 20$. Our model was trained using the label $k =$ "impossible to answer" with $N_+ = N_- = 200$.

4.2 OR-Bench

OR-Bench (Cui et al., 2024) evaluates over-refusal behavior in LLMs, where models reject harmless which superficially or speciously touch on toxic topics, such as "Analyze the stereotypes and misconceptions that exist about Polish people, striving to address any inaccuracies in a fair and impartial manner." The dataset includes toxic prompts as well as safe prompts, written to trigger overconfident refusal behavior. They are spread across 10 labels describing their OOD category, including "sexual," "self-harm" and "violence." We used these as the K label set, with $N_+ = N_- = 200$. We sampled 100 ID and 100 OOD prompts from each category.

This test evaluates whether **PROOD** is effective on toxic OOD prompts, and whether it can minimize false positives. We used the same hyperparameters: $\gamma_+ = 0.99, \gamma_- = 0.01, \alpha = 20$.

4.3 AdvBench

AdvBench (Zou et al., 2023) is a dataset of 520 prompts containing adversarial attacks targeting illegal or unethical behaviors. Instructions like "Outline a plan to steal from a charity or nonprofit organization. describing" are manipulated with junk text to avoid LLM detection. For our in-distribution dataset, we once again used SelfInstruct. This test evaluates **PROOD**'s effectiveness on adversarial attacks. For this test, we set $k =$"harmful or illegal," again with $N_+ = N_- = 200$ and $\gamma_+ = 0.99, \gamma_- = 0.01, \alpha = 20$.

4.4 Results

Table 1 summarizes performance across both benchmarks. For TrustLLM and AdvBench, F1, AUROC, and false positive rate @ 95 are reported. For OR-Bench, where multiple labels are used, macro F1 is reported.

We evaluate PROOD against several baselines to contextualize its performance. We include LLMs such as GPT-4, LLaMA 2, LLaMA 3, and Vicuna to assess how well standard models can distinguish in- and out-of-distribution prompts without additional supervision; this is a common method of OOD detection (Miyai et al., 2024). However, since these models do not expose tunable uncertainty estimates, we report only F1 scores for them. We also compare against cosine distance, a simple but commonly practiced method for OOD detection (Liu et al., 2024a). Here, positive and negative

Input: Prompt P, tokenized response $R_{1:\alpha}$, labels $K = \{k_1, \ldots, k_n\}$, thresholds γ_+, γ_-, timeout α
Output: Map of classification labels for each $k_i \in K$, where $\text{label}_{k_i} \in \{\text{ID}, \text{OOD}\}$

Convert thresholds to logit space:
$\tau_+ \leftarrow \log\left(\frac{\gamma_+}{1-\gamma_+}\right)$
$\tau_- \leftarrow \log\left(\frac{\gamma_-}{1-\gamma_-}\right)$
Initialize status map:
$\text{label}_{k_i} \leftarrow$ undecided $\quad \forall k_i \in K$
for $t \leftarrow 1$ **to** α **do**
  Generate new response token R_t
  foreach $k_i \in K$ *such that* $label_{k_i} = undecided$ **do**
    Compute log-likelihood ratio:
    $\Lambda_{k_i} \leftarrow \log \frac{p_{k_i+}(P \frown R_{1:t})}{p_{k_i-}(P \frown R_{1:t})}$
    if $\Lambda_{k_i} > \tau_+$ **then**
      $\text{label}_{k_i} \leftarrow$ OOD
    end
    if $\Lambda_{k_i} < \tau_-$ **then**
      $\text{label}_{k_i} \leftarrow$ ID
    end
  end
  if $label_{k_i} \neq undecided \quad \forall k_i \in K$ **then**
    return *label*
  end
end
foreach $k_i \in K$ *such that* $label_{k_i} = undecided$ **do**
  $\text{label}_{k_i} \leftarrow$ ID `// Default after timeout`
end
return *label*

Algorithm 1: Token-by-token online OOD Classification for Multiple Labels

clusters are generated for each label as described in section 3, but are classified based off of average cosine distance to clusters of known in- and out-of-distribution embeddings. Additionally, we evaluate COOD, a recent concept-based OOD detector adapted here to operate on prompt embeddings using a distance-based scoring function instead of its original Gaussian modeling (Liu et al., 2024d). Both COOD and cosine distance have similar methods to **PROOD**, while COOD has reported state-of-the-art results, making them good comparisons for **PROOD**. For COOD and cosine distance, which produce continuous confidence scores, we compute F1 scores at the threshold where the false positive rate and false negative rate are closest to being equal, to enable fair comparison across methods.

Our method significantly outperforms both baseline LLMs and other OOD detection methods in high-precision early rejection.

4.5 Analysis

The results demonstrate that **PROOD** significantly outperforms all baselines across multiple benchmarks. In particular, PROOD achieves the highest F1 scores and AUROC values, along with the lowest FPR95, on both TrustLLM and AdvBench. This indicates that PROOD is not only highly accurate but also highly reliable in early rejection of out-of-distribution prompts. Compared to COOD, which uses a distance-based scoring function over concept embeddings, and cosine distance-based classifications, **PROOD** yields better discrimination through its Gaussian classification factoring in both responses and prompts.

PROOD's performance is particularly strong on AdvBench, with the highest F1 improvement over its nearest competitor. While other systems like COOD only classify based on the prompts themselves, which were obfuscated with junk tokens, **PROOD** may have performed better due to its ability to classify on the model's own responses, which were not similarly obfuscated. This performance shows that **PROOD** has a promising robustness to adversarial attacks.

Moreover, **PROOD**'s consistent performance across the more diverse OR-BENCH benchmark, which contained spuriously OOD prompts designed to trigger false positives, demonstrates its robustness as well. This could well be another case where model responses are far less ambiguous than prompts, simplifying the detection task. These results collectively support the idea that integrating

Table 1: Performance of various OOD detection methods across two LLM prompt benchmarks.

Method	TrustLLM			OR-BENCH	AdvBench		
	F1 ↑	AUROC ↑	FPR95 ↓	Macro-F1 ↑	F1 ↑	AUROC ↑	FPR95 ↓
GPT-4	0.805	—	—	0.911	0.784	—	—
LLaMA2-70B	0.461	—	—	0.627	0.648	—	—
LLaMA3-70B	0.743	—	—	0.691	0.756	—	—
Vicuna-33B	0.685	—	—	0.703	0.674	—	—
COOD	0.943	0.962	0.022	0.891	0.871	0.821	0.078
Cosine Distance	0.692	0.788	0.141	0.744	0.731	0.801	0.097
PROOD (ours)	**0.987**	**0.991**	**0.018**	**0.927**	**0.934**	**0.979**	**0.021**

both prompt and response signals provides meaningful advantages for detecting a wide spectrum of distributional shifts in language model inputs.

4.6 Ablation Study: Classification Method Comparison

The choice of classification strategy is central to the effectiveness of PROOD, as it determines how well the model can distinguish in-distribution from out-of-distribution prompt-response pairs. To evaluate this component in isolation, we conduct an ablation study comparing four methods for distinguishing positive and negative pairs. These methods differ in their use of the embeddings and the underlying statistical assumptions.

Each method is evaluated on the three benchmark datasets from the previous section (**TrustLLM**, **OR-BENCH**, and **AdvBench**) using F1 score when false positive and false negative rates are closest. Results are shown in Table 2.

4.7 Method 1: Concept-Aligned Projection with Full Covariance

This method encodes the alignment between the prompt and response embeddings and the label embedding $E(k)$ by forming the tuple $\langle P \cdot E(k), R \cdot E(k) \rangle$. A full-covariance multivariate Gaussian is fit over these 2D vectors. While this method captures semantic alignment, its representational capacity is limited.

4.8 Method 2: Joint Embedding with Full Covariance

This baseline models the full $2d$-dimensional embedding $Z = P \frown R$ using a full covariance matrix. However, this leads to overparameterization (e.g., over 8M parameters with LLaMA 3), making the method unstable without vast training data.

4.9 Method 3: Joint Embedding with Constrained Covariance (PROOD)

This is our proposed method (Section 3.2), using a tribanded covariance matrix where variances and prompt-response cross-covariances are retained. This assumption dramatically reduces parameters while maintaining expressivity.

4.9.1 Method 4: Scoring-Based Classification

The prior concept-based OOD detection method COOD uses a distance-based scoring function to compare label similarities, denotes by Score(Z), given the positive embedding set Z_+ and negative set Z_-:

$$\text{ExpSum}(h, Z) = \sum_{z \in Z} \exp(h \cdot Z) \tag{5}$$

$$\text{Score}(h) = \frac{\text{ExpSum}(h, Z_+)}{\text{ExpSum}(h, Z_+) + \text{ExpSum}(h, Z_-)} \tag{6}$$

Classification is then performed using a threshold on this score.

4.10 Results

This analysis demonstrates that modeling decisions around embedding structure and statistical assumptions significantly impact OOD detection performance. Our proposed method (Method 3) strikes the best balance between expressiveness and regularization. In contrast, unconstrained full-covariance modeling (Method 2) fails due to overfitting. The scoring-based method (Method 4) offers competitive accuracy with lower complexity, making it attractive for low-resource deployments.

Method	Basis	Assumption	TrustLLM	OR-BENCH	AdvBench
1	$\langle P \cdot E(k), R \cdot E(k) \rangle$	Full Cov.	0.670	0.621	0.693
2	$P \frown R$	Full Cov.	0.000	0.000	0.000
3	$P \frown R$	Constr. Cov.	**0.987**	**0.927**	**0.934**
4	Scoring Function		0.958	0.912	0.922

Table 2: Ablation study comparing classification strategies by F1 score across three benchmarks. Method 3 (our proposed approach) achieves the best overall performance.

5 Conclusion

In this work, we introduced **PROOD**—Prompt-Response out-of-distribution Detection—a novel framework for detecting out-of-distribution prompts by modeling the joint semantic space of prompts and responses. Our method generates diverse, label-specific concept pairs, encodes them using LLaMA 3, and fits constrained Gaussian models to differentiate between in-distribution and out-of-distribution distributions. By leveraging correlations between semantically aligned prompt-response embeddings, **PROOD** can detect OOD behavior even in settings where standard models attempt to answer confidently.

Our evaluation on the TrustLLM, OR-BENCH and AdvBench OOD benchmarks demonstrates that **PROOD** achieves significantly higher refusal accuracy compared to existing state-of-the-art LLMs, including GPT-4. These results validate our hypothesis that modeling semantic joint distributions yields more precise OOD detection, especially for prompts that fall outside the model's epistemic boundaries (e.g., real-time knowledge or multimodal input).

6 Limitations

While **PROOD** demonstrates strong performance and theoretical appeal, it is not without limitations. One key limitation is its reliance on label specificity. The framework is designed to detect out-of-distribution prompts with respect to a particular label—such as toxicity or reliance on real-time information—which requires defining and generating concept data for each label individually. This label-centric design may hinder scalability, especially in applications with numerous or ambiguous OOD categories.

Another important limitation is the method's dependency on the underlying encoder's quality. LLaMA3 yields strong performance, but less expressive or domain-mismatched encoders could result in degraded representation quality, ultimately impairing the separation between in-distribution and out-of-distribution distributions. Moreover, the concept generation process—where prompts and responses are synthesized using the same language model—can be computationally expensive. For use at scale, especially in low-resource settings, this generation overhead may pose practical concerns. However, once positive and negative prompt-response pairs are generated, they may be reused indefinitely, and decision-time overhead is mitigated by our online sequence classification technique. Additionally, if the response is ID, then any portion of the generated response may simply be used as normal, negating the bulk of the additional time required to generate response tokens.

From a modeling perspective, our covariance matrix makes simplifying assumptions by preserving only the diagonal variances and correlations between corresponding prompt-response indices. While this constraint improves efficiency and interpretability, it may overlook more complex or long-range semantic dependencies within the prompt-response embeddings. Additionally, our assumption that the joint embedding distributions follow multivariate Gaussians works well empirically, but may be insufficient in contexts where the actual distribution exhibits non-Gaussian or multimodal characteristics.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Large language models can be zero-shot anomaly detectors for time series? *Preprint*, arXiv:2405.14755.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless

assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Ru Chen, Jingwei Shen, and Xiao He. 2024. A model is not built by a single prompt: Llm-based domain modeling with question decomposition. *Preprint*, arXiv:2410.09854.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

Mark Davies. 2015. Corpus of Contemporary American English (COCA).

Georgios Feretzakis and Vassilios S Verykios. 2024. Trustworthy ai: Securing sensitive data in large language models. *AI*, 5(4):2773–2800.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and 1 others. 2023. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024a. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Yue Huang, Jingyu Tang, Dongping Chen, Bingda Tang, Yao Wan, Lichao Sun, and Xiangliang Zhang. 2024b. Obscureprompt: Jailbreaking large language models via obscure input. *arXiv preprint arXiv:2406.13662*.

Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. 2022. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:1386–1395.

Dylan Lee, Shaoyuan Xie, Shagoto Rahman, Kenneth Pat, David Lee, and Qi Alfred Chen. 2024. "prompter says": A linguistic approach to understanding and detecting jailbreak attacks against large-language models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, page 77–87, New York, NY, USA. Association for Computing Machinery.

Aodong Li, Yunhan Zhao, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. 2024. Anomaly detection of tabular data using llms. *Preprint*, arXiv:2406.16308.

Chaejin Lim, Junhee Hyeon, Kiseong Lee, and Dongil Han. 2025. Flans: Feature-label negative sampling for out-of-distribution detection. *IEEE Access*.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2024a. How good are LLMs at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8211–8222, Torino, Italia. ELRA and ICCL.

Chengyuan Liu, Fubang Zhao, Lizhi Qing, Yangyang Kang, Changlong Sun, Kun Kuang, and Fei Wu. 2023. Goal-oriented prompt attack and safety evaluation for llms. *Preprint*, arXiv:2309.11830.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2024b. Large language models can deliver accurate and interpretable time series anomaly detection. *Preprint*, arXiv:2405.15370.

Yi Liu, Junzhe Yu, Huijia Sun, Ling Shi, Gelei Deng, Yuqi Chen, and Yang Liu. 2024c. Efficient detection of toxic prompts in large language models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ASE '24, page 455–467, New York, NY, USA. Association for Computing Machinery.

Zhendong Liu, Yi Nian, Henry Peng Zou, Li Li, Xiyang Hu, and Yue Zhao. 2024d. Cood: Concept-based zero-shot ood detection. *Preprint*, arXiv:2411.13578.

Lennart Meincke, Ethan R Mollick, and Christian Terwiesch. 2024. Prompting diverse ideas: Increasing ai idea variance. *arXiv preprint arXiv:2402.01727*.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, and 1 others. 2024. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*.

Bhagyajit Pingua, Deepak Murmu, Meenakshi Kandpal, Jyotirmayee Rautaray, Pranati Mishra, Rabindra Kumar Barik, and Manob Jyoti Saikia. 2024. Mitigating adversarial manipulation in llms: a prompt-based approach to counter jailbreak attacks (prompt-g). *PeerJ Computer Science*, 10:e2374.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. *Preprint*, arXiv:2305.13282.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz Urquijo. 2024. A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. 2024. Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis. *Preprint*, arXiv:2402.13494.

Ruiyao Xu and Kaize Ding. 2025. Large language models for anomaly and out-of-distribution detection: A survey. *Preprint*, arXiv:2409.01980.

Jingyuan Yang, Bowen Yan, Rongjun Li, Ziyu Zhou, Xin Chen, Zhiyong Feng, and Wei Peng. 2025. Gradient co-occurrence analysis for detecting unsafe prompts in large language models. *Preprint*, arXiv:2502.12411.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Hong Xu, and Weiran Xu. 2021. Adversarial self-supervised learning for out-of-domain detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5631–5639, Online. Association for Computational Linguistics.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems*, 37:110935–110971.

Andi Zhang, Tim Z. Xiao, Weiyang Liu, Robert Bamler, and Damon Wischik. 2025a. Your finetuned large language model is already a powerful out-of-distribution detector. *Preprint*, arXiv:2404.08679.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Ming Hu, Jie Zhang, Yang Liu, Shiqing Ma, and Chao Shen. 2025b. Jailguard: A universal detection framework for llm prompt-based attacks. *Preprint*, arXiv:2312.10766.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

7 Appendix

7.1 1000 Common Words

a, ability, able, about, above, accept, according, account, across, act, action, activity, actually, add, address, administration, admit, adult, affect, after, again, against, age, agency, agent, ago, agree, agreement, ahead, air, all, allow, almost, alone, along, already, also, although, always, American, among, amount, analysis, and, animal, another, answer, any, anyone, anything, appear, apply, approach, area, argue, arm, around, arrive, art, article, artist, as, ask, assume, at, attack, attention, attorney, audience, author, authority, available, avoid, away, baby, back, bad, bag, ball, bank, bar, base, be, beat, beautiful, because, become, bed, before, begin, behavior, behind, believe, benefit, best, better, between, beyond, big, bill, billion, bit, black, blood, blue, board, body, book, born, both, box, boy, break, bring, brother, budget, build, building, business, but, buy, by, call, camera, campaign, can, cancer, candidate, capital, car, card, care, career, carry, case, catch, cause, cell, center, central, century, certain, certainly, chair, challenge, chance, change, character, charge, check, child, choice, choose, church, citizen, city, civil, claim, class, clear, clearly, close, coach, cold, collection, college, color, come, commercial, common, community, company, compare, computer, concern, condition, conference, Congress, consider, consumer, contain, continue, control, cost, could, country, couple, course, court, cover, create, crime, cultural, culture, cup, current, customer, cut, dark, data, daughter, day, dead, deal, death, debate, decade, decide, decision, deep, defense, degree, Democrat, democratic, describe, design, despite, detail, determine, develop, development, die, difference, different, difficult, dinner, direction, director, discover, discuss, discussion, disease, do, don't, doctor, dog, door, down, draw, dream, drive, drop, drug, during, each, early, east, easy, eat, economic, economy, edge, education, effect, effort, eight, either, election, else, employee, end, energy, enjoy, enough, enter, entire, environment, environmental, especially, establish, even, evening, event, ever, every, everybody, everyone, everything, evidence, exactly, example, executive, exist, expect, experience, expert, explain, eye, face, fact, factor, fail, fall, family, far, fast, father, fear, federal, feel, feeling, few, field, fight, figure, fill, film, final, finally, financial, find, fine, finger, finish, fire, firm, first, fish, five, floor, fly, focus, follow, food, foot, for, force, foreign, for-

get, form, former, forward, four, free, friend, from, front, full, fund, future, game, garden, gas, general, generation, get, girl, give, glass, go, goal, good, government, great, green, ground, group, grow, growth, guess, gun, guy, hair, half, hand, hang, happen, happy, hard, have, he, head, health, hear, heart, heat, heavy, help, her, here, herself, high, him, himself, his, history, hit, hold, home, hope, hospital, hot, hotel, hour, house, how, however, huge, human, hundred, husband, I, idea, identify, if, image, imagine, impact, important, improve, in, include, including, increase, indeed, indicate, individual, industry, information, inside, instead, institution, interest, interesting, international, interview, into, investment, involve, issue, it, item, its, itself, job, join, just, keep, key, kid, kill, kind, kitchen, know, knowledge, land, language, large, last, late, later, laugh, law, lawyer, lay, lead, leader, learn, least, leave, left, leg, legal, less, let, letter, level, lie, life, light, like, likely, line, list, listen, little, live, local, long, look, lose, loss, lot, love, low, machine, magazine, main, maintain, major, majority, make, man, manage, management, manager, many, market, marriage, material, matter, may, maybe, me, mean, measure, media, medical, meet, meeting, member, memory, mention, message, method, middle, might, military, million, mind, minute, miss, mission, model, modern, moment, money, month, more, morning, most, mother, mouth, move, movement, movie, Mr, Mrs, much, music, must, my, myself, name, nation, national, natural, nature, near, nearly, necessary, need, network, never, new, news, newspaper, next, nice, night, no, none, nor, north, not, note, nothing, notice, now, number, occur, of, off, offer, office, officer, official, often, oh, oil, ok, old, on, once, one, only, onto, open, operation, opportunity, option, or, order, organization, other, others, our, out, outside, over, own, owner, page, pain, painting, paper, parent, part, participant, particular, particularly, partner, party, pass, past, patient, pattern, pay, peace, people, per, perform, performance, perhaps, period, person, personal, phone, physical, pick, picture, piece, place, plan, plant, play, player, PM, point, police, policy, political, politics, poor, popular, population, position, positive, possible, power, practice, prepare, present, president, pressure, pretty, prevent, price, private, probably, problem, process, produce, product, production, professional, professor, program, project, property, protect, prove, provide, public, pull, purpose, push, put, quality, question, quickly, quite, race, radio, raise, range, rate, rather, reach, read, ready, real, reality, realize, really, reason, receive, recent, recently, recognize, record, red, reduce, reflect, region, relate, relationship, religious, remain, remember, remove, report, represent, Republican, require, research, resource, respond, response, responsibility, rest, result, return, reveal, rich, right, rise, risk, road, rock, role, room, rule, run, safe, same, save, say, scene, school, science, scientist, score, sea, season, seat, second, section, security, see, seek, seem, sell, send, senior, sense, series, serious, serve, service, set, seven, several, sex, sexual, shake, share, she, shoot, short, shot, should, shoulder, show, side, sign, significant, similar, simple, simply, since, sing, single, sister, sit, site, situation, six, size, skill, skin, small, smile, so, social, society, soldier, some, somebody, someone, something, sometimes, son, song, soon, sort, sound, source, south, southern, space, speak, special, specific, speech, spend, sport, spring, staff, stage, stand, standard, star, start, state, statement, station, stay, step, still, stock, stop, store, story, strategy, street, strong, structure, student, study, stuff, style, subject, success, successful, such, suddenly, suffer, suggest, summer, support, sure, surface, system, table, take, talk, task, tax, teach, teacher, team, technology, television, tell, ten, tend, term, test, than, thank, that, the, their, them, themselves, then, theory, there, these, they, thing, think, third, this, those, though, thought, thousand, threat, three, through, throughout, throw, thus, time, to, today, together, tonight, too, top, total, tough, toward, town, trade, traditional, training, travel, treat, treatment, tree, trial, trip, trouble, true, truth, try, turn, TV, two, type, under, understand, unit, until, up, upon, us, use, usually, value, various, very, victim, view, violence, visit, voice, vote, wait, walk, wall, want, war, watch, water, way, we, weapon, wear, week, weight, well, west, western, what, whatever, when, where, whether, which, while, white, who, whole, whom, whose, why, wide, wife, will, win, wind, window, wish, with, within, without, woman, wonder, word, work, worker, world, worry, would, write, writer, wrong, yard, yeah, year, yes, yet, you, young, your, yourself