CED: Comparing Embedding Differences for Detecting Out-of-Distribution and Hallucinated Text

Hakyung Lee[1], **Keon-Hee Park**[2], **Hoyoon Byun**[1], **Jeyoon Yeom**[1], **Jihee Kim**[1]
Gyeong-Moon Park[2*], **Kyungwoo Song**[1*]
[1]Yonsei University, Republic of Korea
[2]Kyung Hee University, Republic of Korea

Abstract

Detecting out-of-distribution (OOD) samples is crucial for ensuring safety and robustness of models deployed in real-world scenarios. While most OOD detection studies focus on fine-tuned models trained on in-distribution (ID) data, detecting OOD in pre-trained models is also important due to computational limits and the widespread use of open-source models. However, pre-trained models often underperform in same domain shift scenarios, as both ID and OOD samples originate from the same domain, leading to high overlap in their embeddings. To address this issue, we propose CED, a training-free OOD detection method that enhances the distinction between ID and OOD samples. We theoretically validate that strategically selected auxiliary and oracle samples improve this separation. On the basis of our theoretical analysis, CED utilizes these specially designed samples to significantly improve the ability of pre-trained models to differentiate ID from OOD samples in text classification and hallucination detection tasks. We verify that CED is a plug-and-play method compatible with various backbone networks like RoBERTa, Llama, and OpenAI Embedding.

1 Introduction

Language models trained on large-scale datasets with extensive parameters, known as pre-trained language models (PLMs), are renowned for their generalization abilities. Consequently, a wide range of research has focused on improving PLM architectures, such as RoBERTa (Liu et al., 2019), Llama (Meta, 2024), and GPT3 (Brown et al., 2020). With the advancement of PLMs, their applications expanded across various fields, demonstrating superior performance on a variety of discriminative and generative tasks, including classification and text completion.

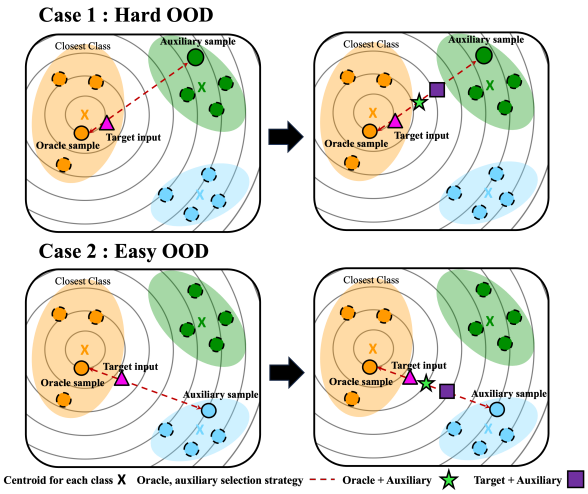


Figure 1: Illustration of CED. Left: Embedding space before CED. Right: Embedding space after applying CED. CED enhances separation of OOD samples in both hard and easy scenarios by leveraging auxiliary and oracle embeddings.

Given the exceptional performance of recent PLMs, there is growing interest in deploying them for real-world applications. Ensuring reliability in these contexts requires the ability to distinguish between in-distribution (ID) and out-of-distribution (OOD) datasets, as test distributions may shift dynamically. This capability becomes especially important in critical sectors like healthcare and customer services, where the stakes are high.

To effectively manage the dynamic nature of real-world data, recent research has focused on fine-tuning PLMs on ID data, and subsequently employing OOD detectors based on the fine-tuned models (Lee et al., 2018; Sun et al., 2022; Lin and Gu, 2023). However, fine-tuning approach faces some limitations. The growing size and complexity of models make fine-tuning increasingly costly and time-consuming. Moreover, the need for constant fine-tuning becomes impractical, especially when new data types or tasks frequently emerge.

*Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14866–14882

These constraints highlight the need for OOD detection methods that can effectively utilize pre-trained models without relying on task-specific fine-tuning.

However, current PLMs struggle with OOD detection, particularly in semantic shift scenarios where ID and OOD datasets share the same domain but have different semantics (Hendrycks et al., 2020; Uppaal et al., 2023; Chen et al., 2023). The performance drop in semantic shift settings occurs because both ID and OOD samples originate from the same domain, resulting in a high overlap in the embeddings.

Therefore, we propose Comparing Embedding Difference (CED), a novel method for OOD detection focusing primarily on classification. CED exploits the distinct relational patterns between ID and OOD samples by concatenating the target input with strategically selected ID samples. This concatenation process enables PLMs to leverage their embedded general knowledge (Hendrycks et al., 2020; Podolskiy et al., 2021), creating a context that amplifies the differences between ID and OOD samples in the embedding space. We show that this approach leads to more accurate OOD detection through a calibrated score derived from embedding differences. As illustrated in Figure 1, CED utilizes the relative differences between perturbed samples in the feature space. We demonstrate the effectiveness and scalability of CED through extensive experiments across diverse datasets and models.

Furthermore, we extend the application of CED to generative tasks, specifically addressing the challenge of hallucination detection. This extension is crucial because generating reliable outputs is necessary for the trustworthy utilization of PLMs in real-world applications. Hallucination detection can be framed as a binary classification problem to determine whether a generated output is factually consistent with the input or contains fabricated information (Su et al., 2024; Ji et al., 2023). This perspective allows us to apply our OOD detection framework directly to the hallucination detection task. To the best of our knowledge, CED is the first method to explore OOD detection for discriminative and generative tasks simultaneously without additional PLM training. We summarize our contributions as follows:

- We derive theoretical findings demonstrating that OOD detection is feasible for certain conditions of auxiliary and oracle datasets.
- Based on theoretical insights, we propose a training-free OOD detection method, CED, for discriminative tasks.
- We extend CED to generative tasks, showing its effectiveness in hallucination detection and its versatility across various language modeling challenges.

2 Related Works

2.1 Types of Data Distribution Shift

Research in data distribution shifts categorizes them mainly into two types: semantic shifts and background shifts (Hsu et al., 2020; Arora et al., 2021). Semantic shifts involve the introduction of entirely new semantic categories that were not part of the ID data. Background shifts, also known as non-semantic shifts, occur when the ID and OOD data share semantic classes but differ in background details or stylistic elements. Prior research shows that pre-trained models generally outperform fine-tuned models in handling non-semantic shifts, demonstrating their robustness in maintaining performance consistency across varied backgrounds (Hendrycks et al., 2020; Uppaal et al., 2023; Chen et al., 2023). In this paper, we focus on demonstrating CED's effectiveness in addressing semantic shifts, an area where pre-trained models typically face challenges.

2.2 OOD Detection Methods

OOD detection methods are essential for ensuring model reliability by identifying inputs that deviate from the training data distribution. These methods typically employ a scoring function to assign OOD scores, with higher scores indicating a greater likelihood of being out-of-distribution (Hendrycks and Gimpel, 2016; Lang et al., 2023). The most common methods include output-based methods, density-based methods (Lin and Xu, 2019), distance-based methods (Zhou et al., 2021) and ensemble-based methods (Baran et al., 2023; Gal and Ghahramani, 2016). Among these approaches, distance-based methods utilize feature embeddings extracted from a model and assume that OOD samples are relatively distant from the ID data. Techniques such as Mahalanobis distance (Lee et al., 2018), K-Nearest Neighbors (Sun et al., 2022), and Flats (Lin and Gu, 2023) are effective but predominantly applied to fine-tuned models. In contrast, our proposed method extends the distance-based

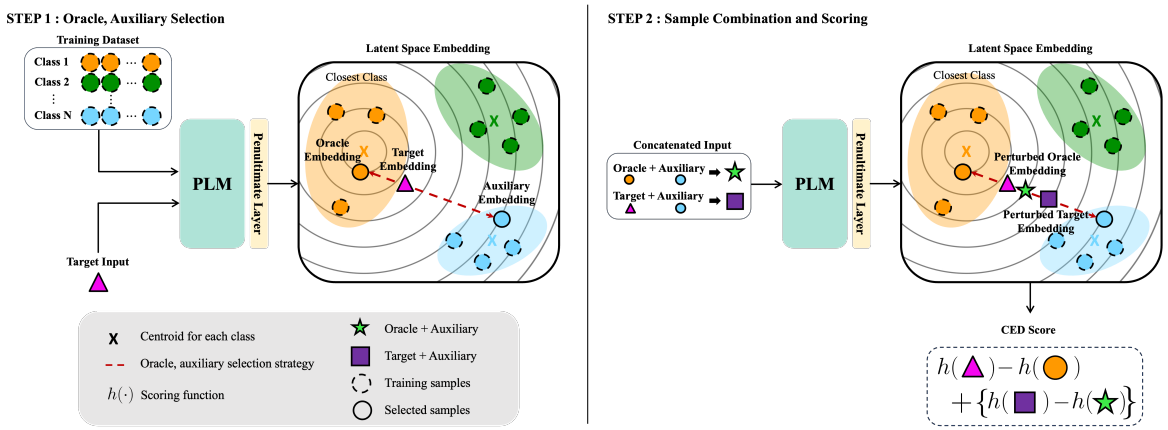


Figure 2: Our proposed CED method for OOD detection. In Step 1, oracle and auxiliary samples are strategically selected from the ID data. In Step 2, the target input is combined with these samples and passed through the PLM to compute the CED score. Red dashed lines indicate the oracle, auxiliary selection strategy validated through theoretical analysis.

approach to work effectively with pre-trained models. By exploiting the relational differences of ID and OOD samples, we enhance OOD detection performance without the necessity for fine-tuning, thus maintaining the pre-trained model's integrity and reducing computational overhead.

2.3 Hallucination Detection

Hallucination detection within large language models (LLMs) has attracted significant attention from researchers as LLMs are widely used across various applications. Liu et al. (2022) introduces a hallucination detection dataset comprising texts derived from perturbed factual content. Kadavath et al. (2022) proposes a self-evaluation method where an LLM is utilized to assess its predictions. Recently, Mündler et al. (2023) aims to detect hallucinations by identifying contradictions between two sampled sentences. Manakul et al. (2023) introduces SelfCheckGPT, which is a black-box approach for detecting hallucinations in LLMs. Zhang et al. (2023) proposes a hallucination detection approach focusing on preceding words and token properties. Further, Su et al. (2024) introduces MIND, which leverages the internal states of LLMs in an unsupervised manner for real-time hallucination detection. In this paper, we propose a training-free method CED that utilizes embeddings with auxiliary samples to detect hallucinations effectively.

3 Method

Our proposed method CED consists of two main steps: 1) Oracle and Auxiliary Sample Selection, and 2) Sample Combination and Scoring. Figure 2 provides an overview of the entire methodology.

3.1 Step 1 : Oracle, Auxiliary Selection

The first step in our method involves the strategic selection of oracle and auxiliary samples from the ID training data. We denote the target input as x_t, which we aim to classify as either ID or OOD. First, we identify the closest class c^* to the target input x_t using Mahalanobis distance, $c^* = \underset{c \in \mathcal{C}}{\operatorname{argmin}} (x_t - \mu_c)^T \Sigma_c^{-1} (x_t - \mu_c)$ where $\mathcal{C}$ is the set of all classes, and μ_c and Σ_c are the mean and covariance of the embeddings for class c, respectively. Let $\mathcal{X}_{c^*}$ be the set of all training samples belonging to class c^*. Within $\mathcal{X}_{c^*}$, we select a set of M oracle samples $\{x_{o_i}\}_{i=1}^M$ that are closest to target input x_t in Euclidean distance,

$$\{x_{o_i}\}_{i=1}^M = \underset{\{x_{o_1},\ldots,x_{o_M}\} \subset \mathcal{X}_{c^*}}{\operatorname{argmin}} \|x_{o_i} - x_t\|_2.$$

Oracle selection identifies samples that are closest to the target input in the embedding space. At this stage, easy OOD samples (those with clearly distinct features from ID samples) may show noticeable differences from their oracle samples, while hard OOD samples (those more similar to ID samples in the embedding space) and ID test samples are likely to appear similar to their respective oracle samples. As hard OOD samples are challenging to distinguish, we utilize CED's auxiliary sample strategy. We hypothesize that when these target and oracle samples are combined with dissimilar auxiliary samples respectively in Step 2, a more distinct pattern will emerge, particularly for hard OOD cases. The contextual embeddings of ID test samples are expected to maintain consistency with their oracle samples, while OOD samples are likely to

show greater divergence after concatenation. This difference in behavior is expected to enhance the effectiveness of OOD detection.

We identify a set of N auxiliary samples, $\{x_{a_j}\}_{j=1}^{N}$, that are oriented in the opposite direction from the oracle embedding relative to the target embedding and are most distant from the target input as follows:

$$\Big\{x_{aj} \mid x_{aj} \in \arg\text{top-N}_{x_{a_j}\in\mathcal{X}} \left\|x_t - x_{a_j}\right\|, \quad \text{s.t.} \quad (x_t - x_{oi})(x_t - x_{a_j}) < 0\Big\}.$$

Our selection process ensures that the auxiliary samples provide contrasting information to the target input, which is crucial for distinguishing x_t from ID data. Figure 2 visualizes the relationships between target inputs, oracle samples, and auxiliary samples in the embedding space. CED captures a comprehensive representation of the target input's relationship to ID data by considering both similarity through close oracle samples and dissimilarity through distant auxiliary samples. This carefully designed selection enables accurate classification of the target input as either ID or OOD, even in challenging scenarios where OOD samples are close to ID data in the embedding space. A detailed case study can be found in Section 4.3. Also, we validate the effectiveness of our selection strategy through theoretical analysis.

3.2 Step 2 : Sample Combination and Scoring

In this step, we concatenate the target input and oracles with selected auxiliary samples to create new input sequences: $\{\tilde{x}_{ta_j}\}_{j=1}^{N}, \{\tilde{x}_{o_i a_j}\}_{i=1,j=1}^{M,N}$. These concatenated inputs are processed through the PLM to obtain feature representations $f(\cdot)$, leveraging the inherent knowledge of PLMs (Radford et al., 2019; Kenton and Toutanova, 2019; Brown et al., 2020).

We introduce the CED score to quantify the distinction between ID and OOD data:

$$\text{CED Score}(x_t) = \frac{1}{M \cdot N} \sum_{j=1}^{M} \sum_{n=1}^{N} \Big(h(f(x_t)) - \alpha h(f(x_{o_i})) + \beta \left(h(f(\tilde{x}_{ta_j})) - h(f(\tilde{x}_{o_i a_j})) \right) \Big).$$

Here, $h(\cdot)$ is a scoring function which can be any distance-based method such as Mahalanobis distance or KNN distance, and α, β are scaling factors. The CED score consists of two components. The first term $h(f(x_t)) - h(f(x_{o_i}))$ measures the direct similarity between target input and the oracle samples. For ID test samples, this difference is expected to be small, as the target input is likely to belong to the same class as the oracle samples. The second term $h(f(\tilde{x}_{ta_j})) - h(f(\tilde{x}_{o_i a_j}))$ captures relational differences introduced by auxiliary samples. For ID test samples, the similarity between the concatenated target-auxiliary inputs and oracle-auxiliary inputs remains high, leading to small differences. In contrast, OOD samples exhibit larger differences because OOD samples have inherently different characteristics from oracle samples, leading to more pronounced discrepancies in the combined embeddings. By averaging scores across multiple oracle and auxiliary samples, CED robustly captures relational patterns. Moreover, the plug-and-play nature of CED offers flexibility in choosing the most suitable distance metric for various applications.

When using Mahalanobis distance (MD) as the scoring function, we measure the distance of perturbed samples relative to the original class c^*. Instead of finding the minimum distance class for each perturbation, we consistently evaluate the distance to class c^*, ensuring alignment with the initial class and providing a stable measurement of relational differences. As shown in Figure 2, the concentric circles represent the MD from the closest class (Class 1 in orange), with greater distances indicating higher likelihood of being OOD. Perturbed OOD targets are positioned further from the center, showing their dissimilarity to ID samples.

3.3 Theoretical Analysis

In this section, we theoretically validate the effectiveness of strategic oracle and auxiliary selection in CED method. This selection enables CED to effectively identify hard OOD instances that are difficult to distinguish from the ID dataset. Further details are provided in Appendix C.

Definition 1 (Hard OOD samples)**.** *A hard OOD sample, x_t, refers to an OOD sample where $h(f(x_t))$ is smaller than that of oracle samples, $h(f(x_o))$. In other words, $\exists x_o$ such that $h(f(x_o)) > h(f(x_t))$.*

Theorem 1 states that oracle and auxiliary samples satisfying certain conditions, with the closest oracle samples and auxiliary samples that are the farthest in the opposite direction, enhance the performance on hard OOD datasets.

Theorem 1. *Let $h(\cdot)$ and $f(x) = w^T x + b$ represent the Mahalanobis distance-based OOD score function and a linear function, respectively. Where $w, x \in \mathbb{R}^d$ and $b \in \mathbb{R}$, we define $f(x_t) - \mu_{ID} = \epsilon$, $f(x_t) - f(x_o) = d_{to}$, $f(x_t) - f(x_a) = d_{ta}$, μ_{ID} as the closest mean of the class embeddings with $f(x_t)$. Given hard OOD target x_t, and $\epsilon > (2 + \sqrt{2})d_{to}$, the* CED Score$(x_t)$ *is greater than $h(f(x_t))$, when an auxiliary sample satisfies condition:*

$$d_{ta} < 0 \text{ and } d_{ta} < \frac{(3 + 2\sqrt{2} + \lambda^2 - 6\lambda - 2\sqrt{2}\lambda)d_{to}}{2\lambda(\lambda - 1)},$$

where $0 < \lambda < 1$ and d_{to} is a positive value close to zero.

3.4 Toy Experiments

To support our theoretical analysis, we conduct a toy experiment using simulated data sampled from a 1-dimensional Gaussian distribution. Figure 3 illustrates this experiment, which shows both scenarios when oracle embedding is located positively and negatively relative to the target. In this experiment, we focus on a hard OOD sample close to μ_{ID} and select an auxiliary set that specifies the oracle and ensures a positive CED score. Consistent with our theoretical analysis, we observe that auxiliary samples located in the opposite direction from the oracle, with greater distances, lead to higher CED scores. The farthest auxiliary samples in opposite directions consistently improve OOD scores, regardless of the oracle sample location. Further validation with toy experiments on 2-dimensional data is provided in Appendix D.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our proposed method on two types of tasks: four dialogue intent classification datasets and two news topic categorization datasets, both commonly used in OOD detection research. Detailed descriptions of the datasets can be found in Appendix A.

Models and Metrics We evaluate our method using three backbone models: RoBERTa$_{\text{base}}$ (Liu et al., 2019), Llama 3-8B (Meta, 2024), and OpenAI text-embedding-3-small (OpenAI, 2024). We use AUROC and FPR95 metrics, following prior works (Liu et al., 2020).

Baselines We compare our approach with several OOD detection methods, focusing primarily on three distance-based methods: MD$_{\text{pre}}$ (Lee et al., 2018), KNN$_{\text{pre}}$ (Sun et al., 2022) , and Flats$_{\text{pre}}$ (Lin and Gu, 2023). We include Cosine$_{\text{pre}}$ (Zhou et al., 2021), LOF$_{\text{pre}}$ (Lin and Xu, 2019), and MC$_{\text{pre}}$ (Gal and Ghahramani, 2016) baselines for experiments on RoBERTa$_{\text{base}}$.We adapt all methods to use embeddings from pre-trained models. CED is integrated as a plug-in approach to the three main distance-based methods.

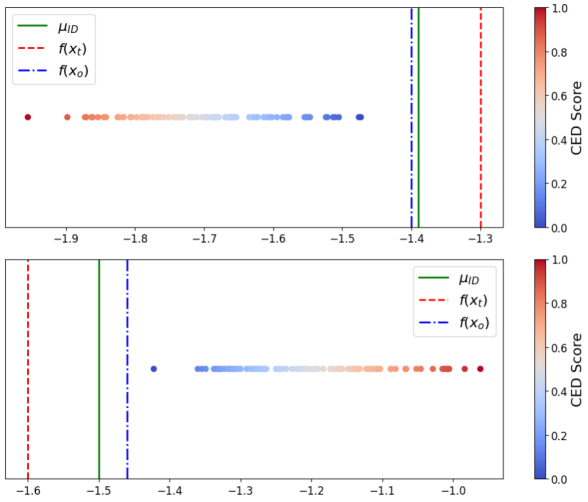


Figure 3: Toy experiment on 1D dataset. Auxiliary samples farther from the target in the opposite direction of the oracle yield higher CED scores (red), consistently improving OOD detection regardless of oracle position.

Hyperparameters We determined the scaling parameters α, β through hyperparameter search within the range of $[0, 1]$, selecting best values for each method. Across all datasets, we use $M = 3$ oracle samples and $N = 5$ auxiliary samples. We conduct sensitivity analysis on both the scaling parameters and the number of samples in Appendix E and Section 4.5. For the KNN approach, we set $k = 10$, following (Chen et al., 2023; Lin and Gu, 2023).

4.2 Comparison with OOD baselines

As shown in Table 1, integrating CED with baseline methods consistently enhances OOD detection performance across various models and datasets. For the Snips dataset, CED led to significant reductions in FPR95 across different models. For example, KNN$_{\text{pre}}$+CED reduced FPR95 from 0.6203 to 0.3743 with RoBERTa and from 0.6524 to 0.3797 with Llama3-8B, demonstrating a substantial improvement. These consistent improvements across multiple datasets and models underscore the effectiveness of CED in enhancing OOD detection performance.

We observed that despite Llama being a LLM,

Model	Methods	CLINC150	ROSTD	Banking77	Snips	NC	AGNews	**Avg.**
		FPR95↓	FPR95↓	FPR95↓	FPR95↓	FPR95↓	FPR95↓	FPR95↓
RoBERTa	$\text{Cosine}_{\text{pre}}$	0.7620	0.9615	0.9444	0.7701	0.9284	0.9511	0.8863
	LOF_{pre}	0.7420	0.6262	0.8972	0.3743	0.8811	0.8103	0.7219
	MC_{pre}	0.9440	0.9278	0.9491	0.9251	0.9490	0.9639	0.9432
	MD_{pre}	0.7330	0.3133	0.9306	0.6364	0.9014	0.8794	0.7324
	MD_{pre} + CED	0.6570	0.1638	0.8843	0.5615	0.8861	0.8194	0.6620
	KNN_{pre}	0.7070	0.1851	0.8935	0.6203	0.8456	0.8189	0.6784
	KNN_{pre} + CED	**0.6450**	**0.0994**	**0.8407**	0.3743	**0.8285**	**0.7575**	**0.5909**
	$\text{Flats}_{\text{pre}}$	0.7460	0.1696	0.9213	0.5348	0.8878	0.7700	0.6716
	$\text{Flats}_{\text{pre}}$ + CED	0.6880	0.1097	0.8852	**0.3529**	0.8753	0.7664	0.6129
Llama3-8B	MD_{pre}	0.8310	0.3398	0.9278	0.5936	0.9198	0.8592	0.7452
	MD_{pre} + CED	0.7450	0.2039	0.9139	0.5294	0.8980	0.7883	0.6797
	KNN_{pre}	0.8070	0.3039	0.9241	0.6524	0.8770	0.8319	0.7327
	KNN_{pre} + CED	**0.6270**	0.1816	**0.8213**	**0.3797**	**0.8479**	**0.7256**	**0.5971**
	$\text{Flats}_{\text{pre}}$	0.8010	0.3003	0.9454	0.5722	0.8987	0.7911	0.7181
	$\text{Flats}_{\text{pre}}$ + CED	0.6470	**0.1641**	0.8769	0.3850	0.8764	0.7789	0.6213
OpenAI text-embedding-3 small	MD_{pre}	0.1840	0.0061	0.5972	0.1070	0.8553	0.6817	0.4052
	MD_{pre} + CED	0.1670	0.0045	0.5852	0.0963	0.8483	0.6747	0.3960
	KNN_{pre}	0.0860	0.0016	0.2907	0.0321	0.5102	**0.5225**	0.2405
	KNN_{pre} + CED	**0.0690**	**0.0016**	**0.2694**	**0.0321**	**0.5064**	0.5289	**0.2298**
	$\text{Flats}_{\text{pre}}$	0.1830	0.0074	0.4685	0.0749	0.7356	0.6217	0.3485
	$\text{Flats}_{\text{pre}}$ + CED	0.1560	0.0061	0.4509	0.0642	0.7276	0.6344	0.3399

Table 1: FPR95 performance of CED across different datasets. The best results for each model is shown in **bold**.

its performance in some cases does not surpass that of RoBERTa. In contrast, the OpenAI text embedding model, which is a model optimized for embeddings, showed best performance. Llama's lower performance may be due to our use of last token embeddings for OOD detection, following prior research (Liu et al., 2024), which might not fully capture the complex contextual information inherent in a decoder-only model.

4.3 Case Study of Sample Interactions

To better understand the influence of each sample type, we conducted a qualitative analysis using the ROSTD dataset with the RoBERTa model, as detailed in Table 2. In the absence of a classification head in pre-trained models, we identified the closest class by calculating the smallest Mahalanobis distance from the embeddings.

When an OOD query ("Give me the current town") is paired with an auxiliary sample about an alarm, the closest class shifts from "Weather/Find" to "Alarm/Snooze Alarm". This shift significantly diverges from the behavior observed when an oracle sample is combined with the same auxiliary, where the classification remains "Weather/Find". In contrast, when an ID sample is combined with an auxiliary sample, the resulting classification closely aligns with that of an oracle sample paired with the same auxiliary. This consistent behavior between ID + auxiliary and oracle + auxiliary pairs sharply contrasts with the differing behavior seen in OOD + auxiliary and oracle + auxiliary pairs, which is a crucial element in distinguishing OOD samples from ID samples.

Figure 4 quantitatively demonstrates CED's effectiveness, complementing the qualitative examples. Normalized OOD scores show that CED enhances the distinction between ID and OOD samples. For example, without CED, the OOD score reaches 3.23 times the ID score, whereas with CED, it amplifies to 5.74 times. This significant increase underscores CED's enhanced capability in OOD detection.

4.4 Ablation Study

In Table 3, we conduct an ablation study on the AGNews dataset using RoBERTa to assess the impact of our sample selection strategy in CED. Random oracle selection significantly lowers performance, highlighting the importance of strategic oracle choice. Randomizing both oracle and auxiliary selections leads to the poorest results, emphasizing

Sample Type	Query Example	Closest Class
OOD	Give me the current town.	Weather/Find
OOD + Auxiliary	Give me the current town. Snooze my alarm for fifteen minutes.	**Alarm/Snooze Alarm**
Oracle + Auxiliary	Give me the current weather in provo. Snooze my alarm for fifteen minutes.	**Weather/Find**
ID	What is the chance of severe weather?	Weather/Find
ID + Auxiliary	What is the chance of severe weather? Show all of today's alarms.	Weather/Find
Oracle + Auxiliary	What is the chance of snow? Show all of today's alarms.	Weather/Find

Table 2: Examples of test samples and selected oracle, auxiliary samples from ROSTD dataset.

Method	AUC↑	FPR↓
Rand. Oracle	0.6695	0.8442
Rand. Aux.	0.7334	0.7667
Rand. Oracle & Rand. Aux.	0.6672	0.8500
In-class Rand. Oracle	0.6693	0.8436
In-class Rand. Oracle & Rand. Aux.	0.6670	0.8489
Position Permutation	0.7351	0.7603
CED	**0.7357**	**0.7600**

Table 3: Ablation study for AG-News Dataset. Rand. refers to random, Aux. refers to auxiliary.

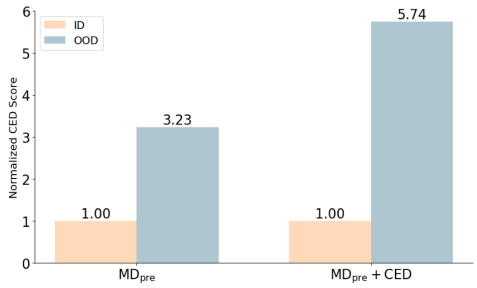


Figure 4: Comparison of MD scores for ID and OOD datasets on ROSTD. Unlike the baseline, CED enhances the difference between OOD and ID scores; therefore, CED shows superior performance.

the need for careful selection of both. In-class random oracle selection, where oracles are randomly chosen within the target's closest class, also underperforms. This suggests that class similarity alone is insufficient, possibly due to poorly clustered embeddings in pre-trained models. Position permutation, whether the auxiliary is placed before or after the target sample, has minimal impact, indicating that sample order is less critical. These findings emphasize the sample selection procedure for CED's effectiveness.

4.5 Sensitivity Analysis

Figure 5 shows the sensitivity analysis of KNN_{pre}+CED applied to RoBERTa, with respect to the number of oracle (M) and auxiliary (N) samples on the Snips dataset. The gray surface represents the baseline performance of KNN_{pre} without CED, which remains constant across all M and N values. Both AUROC and FPR95 plots demonstrate improvement as M and N increase, with peak performance at larger values. The red surface, representing KNN_{pre}+CED, consistently outperforms the baseline KNN_{pre} (gray surface) across all parameter combinations. These results indicate that CED effectively leverages additional contextual information through oracle and auxiliary samples to enhance OOD detection.

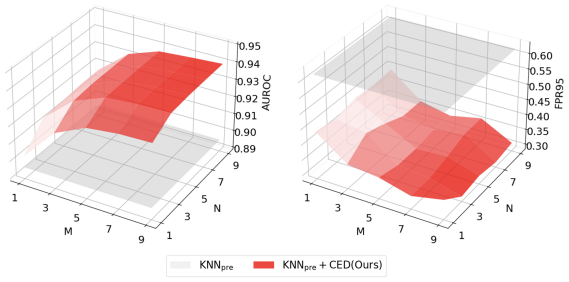


Figure 5: Performance analysis of KNN_{pre}+CED with varying oracle auxiliary parameters M and N on Snips dataset. The surface plots show AUROC (left) and FPR95 (right) scores.

4.6 Hallucination Detection

We extend CED to generative tasks, focusing on detecting hallucinations in text generated by LLMs. Hallucinations occur when a model generates information that is not grounded in the input data or factual reality, which can be approached as a classification problem. By framing hallucination detection as an OOD detection task, we demonstrate that CED can be effectively adapted to mitigate these issues, extending its utility beyond discriminative tasks to generative contexts as well.

	Sentence Level AUC↑			Passage Level AUC↑			Sentence Level Corr↑			Passage Level Corr↑		
Baselines	GPT-J	Llama2	OPT-7B	GPT-J	Llama2	OPT-7B	GPT-J	Llama2	OPT-7B	GPT-J	Llama2	OPT-7B
PE-max	0.7497	0.6851	0.7263	0.8875	0.8400	0.8851	0.0839	0.2032	0.0573	0.2660	0.2561	0.2180
PE-min	0.7044	0.5878	0.7228	0.7595	0.7587	0.8075	-0.0316	0.0152	0.0601	-0.0645	0.0928	0.0823
PP-max	0.7074	0.5872	0.7302	0.7585	0.7543	0.8100	-0.0467	0.0514	0.0732	-0.0672	0.0837	0.0943
PP-min	0.7413	0.6025	0.7121	0.8526	0.7969	0.8384	0.2057	0.1842	0.1808	0.2454	0.2231	0.1191
SCG-MQAG	0.7873	0.6401	0.7593	0.8827	0.8196	0.8594	0.2306	0.1822	0.2151	0.3145	0.2278	0.2993
SCG-NG	0.7549	0.5365	0.7490	0.8579	0.7155	0.8340	0.1770	0.0590	0.1167	0.2222	0.0785	0.1021
SCG-BS	0.7424	0.6178	0.6594	0.8165	0.7631	0.7597	0.0745	0.1268	-0.0563	0.1288	0.1447	-0.0730
SCG-NLI	0.8680	0.7644	0.8103	0.9384	0.8897	0.9096	0.4087	0.3809	0.3312	0.4217	0.4389	0.4091
GPT4-HDM	0.7843	0.6583	0.7972	0.9183	0.8265	0.9196	0.1096	0.0086	0.1678	0.2539	0.1876	0.2372
CED	**0.8816**	**0.7988**	**0.9097**	**0.9661**	**0.9359**	**0.9407**	**0.4789**	**0.4736**	**0.5135**	**0.5397**	**0.5124**	**0.4848**

Table 4: Hallucination detection task for diverse datasets and evaluation metrics. For generated text from GPT-J, Llama2, and OPT-7B, our CED methods consistently show superior performance.

4.6.1 Experimental Settings

We used the HELM benchmark (Su et al., 2024) to detect hallucinations in text generated by LLMs, specifically GPT-J (Wang and Komatsuzaki, 2021), Llama2-7B (Touvron et al., 2023), and OPT-7B (Zhang et al., 2022). HELM involves prompt-based text generation using articles from the WikiText-103 corpus (Merity et al., 2016), with generated sentences manually labeled as hallucinatory or non-hallucinatory. We evaluate performance using Area Under the Curve (AUC) and Pearson correlation coefficient, measured per sentence and passage. CED is implemented similarly to the classification task, with oracle and auxiliary samples selected at the paragraph level using the Wiki articles (Su et al., 2024). We compare the performance of CED against various training-free baselines. Detailed experimental settings are provided in Appendix B.

4.6.2 Comparison with other Baselines

As shown in Table 4, CED achieves the highest sentence-level and passage-level AUC scores, as well as the highest correlation coefficients. Compared to the best-performing baseline, SCG-NLI, CED shows significant improvements in all metrics. Notably, CED demonstrates substantially better performance compared to PP and PE baselines, which rely on logit values for hallucination detection. This suggests that leveraging contextual embeddings is significantly more effective than using logit-based methods. By strategically combining oracle and auxiliary samples, CED enhances the distinction between hallucinatory and non-hallucinatory text, leading to superior performance compared to other training-free methods. Results highlight the importance of exploiting the intrinsic knowledge in pre-trained models and the relational patterns in the embeddings for accurate hallucination detection, even without additional training. Appendix F.2 provides additional qualitative analysis for the hallucination task.

5 Conclusion

With the recent advances and wide application of PLMs, the reliability of PLM on diverse tasks is crucial. However, current PLMs are vulnerable to detecting OOD classes or hallucinated text. In this study, we propose a novel OOD detection method, CED, that effectively harnesses the inherent relational patterns in pre-trained models. First, we present theoretical findings that certain combinations of auxiliary and oracle samples enhance OOD detection performance. Especially, we verify that CED, with a certain sample selection strategy, captures the hard OOD dataset. Second, we implement these theoretical findings into a training-free method called CED. Third, we extensively verify the superiority of CED on both discriminative and generative tasks across diverse datasets. Our method is plug-and-play and compatible with various OOD detection methods and PLM backbones.

Limitations

In this study, we propose a new training-free method for OOD detection in both discriminative and generative tasks. To validate the method, we present extensive experimental results on diverse tasks, datasets, and backbones. However, access to GPT-4o embeddings was not available, so we could not provide compatibility results between CED and GPT-4o. Additionally, further valida-

tion with real-world datasets from various domains, such as medical and engineering, is required. There is also some potential risk in this research, as we validated the effectiveness of our model only on English datasets. Further analysis of diverse language datasets is necessary for future work.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00457216, 2022R1A4A3033874).

References

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701.

Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. 2023. Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 119–129.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Sishuo Chen, Wenkai Yang, Xiaohan Bi, and Xu Sun. 2023. Fine-tuning deteriorates general textual out-of-distribution detection by distorting task-agnostic features. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 564–579.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Hao Lang, Yinhe Zheng, Yixuan Li, SUN Jian, Fei Huang, and Yongbin Li. 2023. A survey on out-of-distribution detection in nlp. *Transactions on Machine Learning Research*.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Heeyoung Lee, Hoyoon Byun, Changdae Oh, JinYeong Bak, and Kyungwoo Song. 2024. Perturb-and-compare approach for detecting out-of-distribution samples in constrained access environments. *arXiv preprint arXiv:2408.10107*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Haowei Lin and Yuntian Gu. 2023. Flats: Principled out-of-distribution detection with feature-based likelihood ratio score. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8956–8963.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2024. 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.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Rishabh Misra and Jigyasa Grover. 2021. Sculpting data for ml: The first act of machine learning. *University of California San Diego: La Jolla, CA, USA*, page 158.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2024. Embeddings. https://platform.openai.com/docs/guides/embeddings.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12813–12832.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.

A Details of Dataset

We conducted experiments on Banking77, CLINC150, ROSTD, SNIPS, AGNews, News Categories and HELM dataset. Details of each dataset are as follows:

Banking77 (Casanueva et al., 2020) is a fine-grained intent classification dataset in the banking domain, consisting of 77 intent classes, with 50 classes used as ID classes and the remaining 22 as OOD classes. Banking77 includes 9,003 user queries for training, 1,000 for validation, and 3,080 for testing.

CLINC150 (Larson et al., 2019) is designed for OOD intent detection, consisting of 150 intent classes. CLINC150 contains 15,000 training instances, 3,000 validation instances, and 4,500 testing instances of ID data, respectively. Additionally, CLINC150 includes 1,000 OOD test instances.

ROSTD (Gangal et al., 2020) is a large-scale intent classification dataset consisting of 12 intent classes. ROSTD contains 30,521, 4,181, and 8,621 samples for the training, validation, and testing, respectively.

SNIPS (Coucke et al., 2018) contains annotated utterances from diverse domains. SNIPS includes 7 intent classes, with 5 classes used as ID classes and the remaining 2 classes as OOD classes. We obtained 9,361 training samples, 500 validation samples, 513 testing samples, and 187 OOD testing samples.

AGNews (Del Corso et al., 2005) consists of 4 classes such as "World", "Sports", "Business", and "Sci/Tech". We split the AGNews dataset into 115,778 samples for training, 3,994 samples for validation, and 3,993 samples for testing. For the OOD detection, we used the 3,600 samples as OOD test samples.

News Categories (NC) (Misra and Grover, 2021) is one of the largest news datasets. We leveraged 5 classes from the News Category as ID data, with 68,859 samples used for training, 8,617 samples for validation, and 8,684 samples for testing. We used the remaining classes as OOD data.

HELM Dataset (Su et al., 2024) The number of sentences generated and the percentage of sentences labeled as hallucinatory/non-hallucinatory for each model are presented below: GPT-J: Generated 572 sentences from 208 paragraphs (172 hallucinations, 400 non-hallucinations). Llama2: Generated 565 sentences from 207 paragraphs (243 hallucinations, 322 non-hallucinations). OPT-7B: Generated 566 sentences from 201 paragraphs (181 hallucinations, 385 non-hallucinations).

B Experimental Details for Hallucination Detection

The experimental settings for hallucination detection are adopted from (Su et al., 2024). Applying CED to the hallucination detection task follows the same procedure as for classification tasks, with the Mahalanobis distance (MD) score used as the distance method. There are two key difference from the classification tasks. First, instead of using the last token embedding, we used the last layer mean token embedding, following the setup from (Su et al., 2024). Second, the oracle and auxiliary samples were selected not from the ID train data, but from the Wiki articles (Su et al., 2024). This was done to avoid the risk of introducing hallucinations through concatenation, as the training data may contain hallucinations.

For experiments, we employed three LLMs, GPT-J (Wang and Komatsuzaki, 2021), Llama2-7B (Touvron et al., 2023), and OPT-7B (Zhang et al., 2022). We compared the performance of hallucination detection using the following training-free baselines, ensuring a fair comparison with CED:

Predictive Probability (PP) (Manakul et al., 2023) Detects hallucinations based on the probability of tokens generated by the language model.

Predictive Entropy (PE) (Kadavath et al., 2022) Detects hallucinations by evaluating the uncertainty in the output distribution of the language model.

SelfCheckGPT (SCG) (Manakul et al., 2023) Detects hallucinations based on the consistency of the model's responses across similar prompts.

GPT4-HDM (Li et al., 2023) Uses GPT-4 to evaluate the outputs of other language models to determine if hallucination is present.

C Proofs

C.1 Proof of Proposition

Proposition 1 (OOD score function for concatenated text with attention mechanism)**.** *Given two sequences $S_1 \in \mathbb{R}^{n\times d}$ and $S_2 \in \mathbb{R}^{m\times d}$, the attention mechanism applied to the concatenated sequence $\mathcal{S} = Concat(S_1, S_2) \in \mathbb{R}^{(n+m)\times d}$ can be interpreted as a linear interpolation of the attention mechanisms applied to each sequence separately. Consider $\tilde{x}_{ta} = \lambda x_t + (1-\lambda)x_a$ where a target sentence $x_t \in S_1$ and an auxiliary sentence $x_a \in S_2$*

and $\lambda \in (0,1)$. This $\tilde{x}_{ta}$ represents a mixed sample formed by a linear interpolation of x_t from S_1 and x_a from S_2. Additionally, let a pre-trained model $f(\cdot)$ and base OOD score function $h(\cdot)$ be twice-differentiable functions. The base OOD score function of the mixed sample, $h(f(\tilde{x}_{ta}))$, can be written as:

$$h(f(\tilde{x}_{ta})) = h(f(x_t)) + \sum_{l=1}^{3} \omega_l(x_t, x_a) + \varphi_t(\lambda)(\lambda-1)^2,$$

where $\lim_{\lambda\to 1} \varphi_t(\lambda) = 0$, and $\omega_l(x_t, x_a)$ are defined as:

$$\begin{aligned}
\omega_1(x_t, x_a) &= (\lambda-1)(x_t - x_a)^T f'(x_t) h'(f(x_t)),\\
\omega_2(x_t, x_a) &= \frac{(\lambda-1)^2}{2}(x_t - x_a)^T f''(x_t)(x_t - x_a)\\
&\quad \cdot h'(f(x_t)),\\
\omega_3(x_t, x_a) &= \frac{(\lambda-1)^2}{2}(x_t - x_a)^T f'(x_t)\\
&\quad \cdot (x_t - x_a)^T f'(x_t) h''(f(x_t)).
\end{aligned}$$

Proof of Proposition. We will first consider the self-attention mechanism for the concatenated sequence $\mathcal{S}$:

$$\begin{aligned}
&\text{Attn}(\mathcal{S}\mathbf{W_q}, \mathcal{S}\mathbf{W_k}, \mathcal{S}\mathbf{W_v})\\
&\quad = \text{softmax}(\mathcal{S}\mathbf{W_q} \cdot (\mathcal{S}\mathbf{W_k})^T)\mathcal{S}\mathbf{W_v}\\
&\quad = \text{softmax}(\text{Concat}(S_1\mathbf{W_q}, S_2\mathbf{W_q}) \cdot\\
&\qquad \text{Concat}(S_1\mathbf{W_k}, S_2\mathbf{W_k})^T) \begin{bmatrix} S_1\mathbf{W_v} \\ S_2\mathbf{W_v} \end{bmatrix}\\
&\quad = \begin{bmatrix} \lambda(S_1)S_1\mathbf{W_v} + (1-\lambda(S_1))S_2\mathbf{W_v} \\ (1-\lambda(S_2))S_1\mathbf{W_v} + \lambda(S_2)S_2\mathbf{W_v} \end{bmatrix},
\end{aligned}$$

where $\lambda(S_1)$ and $\lambda(S_2)$ represent the summation of normalized attention weights on the attention of sequences S_1 and S_2, respectively. These are defined as:

$$\begin{aligned}
\lambda(S_1) &= \frac{\sum \exp(\mathbf{A}_{11})}{\sum \exp(\mathbf{A}_{11}) + \sum \exp(\mathbf{A}_{12})},\\
\lambda(S_2) &= \frac{\sum \exp(\mathbf{A}_{22})}{\sum \exp(\mathbf{A}_{21}) + \sum \exp(\mathbf{A}_{22})},\\
\text{where } \mathbf{A}_{ij} &= S_i \mathbf{W_q} \mathbf{W_k}^T S_j^T, \text{ for } i,j \in \{1,2\}
\end{aligned}$$

Next, we consider the base OOD score function for the mixed sample $\tilde{x}_{ta}$. Let $\psi_t(\lambda) = h(f(\tilde{x}_{ta}))$, which is a modified function of $h(f(\tilde{x}_{ta}))$ with λ as an input. Following the approach proposed in (Lee et al., 2024), we use a second-order Taylor approximation to approximate $\psi_t(\lambda)$:

$$\begin{aligned}
\psi_t(\lambda) &= \psi_t(1) + \psi_t'(1)(\lambda-1)\\
&\quad + \frac{1}{2}\psi_t''(1)(\lambda-1)^2 + \varphi_t(\lambda)(\lambda-1)^2,
\end{aligned}$$

where $\lim_{\lambda\to 1} \varphi_t(\lambda) = 0$.

First, we find $\psi_t'(\lambda)$:

$$\begin{aligned}
\psi_t'(\lambda) &= \frac{\partial \tilde{x}_{ta}}{\partial \lambda} \frac{\partial f(\tilde{x}_{ta})}{\partial \tilde{x}_{ta}} \frac{\partial h(f(\tilde{x}_{ta}))}{\partial f(\tilde{x}_{ta})}\\
&= (x_t - x_a)^T f'(\tilde{x}_{ta}) h'(f(\tilde{x}_{ta})).
\end{aligned}$$

Next, we find $\psi_t''(\lambda)$:

$$\begin{aligned}
\psi_t''(\lambda) &= (x_t - x_a)^T f''(\tilde{x}_{ta})(x_t - x_a) h'(f(\tilde{x}_{ta}))\\
&\quad + (x_t - x_a)^T f'(\tilde{x}_{ta})\\
&\quad \cdot (x_t - x_a)^T f'(\tilde{x}_{ta}) h''(f(\tilde{x}_{ta})).
\end{aligned}$$

When $\lambda = 1$:

$$\begin{aligned}
\psi_t'(1) &= (x_t - x_a)^T f'(x_t) h'(f(x_t)),\\
\psi_t''(1) &= (x_t - x_a)^T f''(x_t)(x_t - x_a) h'(f(x_t))\\
&\quad + (x_t - x_a)^T f'(x_t)\\
&\quad \cdot (x_t - x_a)^T f'(x_t) h''(f(x_t)).
\end{aligned}$$

Combining these results, we derive the desired equation:

$$h(f(\tilde{x}_{ta})) = h(f(x_t)) + \sum_{l=1}^{3} \omega_l(x_t, x_a) + \varphi_t(\lambda)(\lambda-1)^2.$$

□

C.2 Proof of Theorem 1

Theorem 1. *Let $h(\cdot)$ and $f(x) = w^T x + b$ represent the Mahalanobis distance-based OOD score function and a linear function, respectively. Where $w, x \in \mathbb{R}^d$ and $b \in \mathbb{R}$. We define $f(x_t) - \mu_{ID} = \epsilon$, $f(x_t) - f(x_o) = d_{to}$, $f(x_t) - f(x_a) = d_{ta}$, μ_{ID} as the closest mean of the class embeddings with $f(x_t)$. Given hard OOD target x_t, and $\epsilon > (2+\sqrt{2})d_{to}$, the* CED Score$(x_t)$ *is greater than $h(f(x_t))$, when an auxiliary sample satisfies condition:*

$$d_{ta} < 0 \text{ and}$$

$$d_{ta} < \frac{(3 + 2\sqrt{2} + \lambda^2 - 6\lambda - 2\sqrt{2}\lambda) d_{to}}{2\lambda(\lambda-1)},$$

where constant $0 < \lambda < 1$ and d_{to} is a positive value close to zero.

Proof of Theorem 1. Under the assumption that $f(x)$ is a linear model $w^T x$, where $w,\ x \in \mathbb{R}^d$, Mahalanobis score $h(x)$ can be expressed as $h(f(x)) = (f(x) - \mu_{\text{ID}})^T \Sigma_{\text{ID}}^{-1}(f(x) - \mu_{\text{ID}}) = \frac{1}{\sigma_{\text{ID}}}(f(x) - \mu_{\text{ID}})^2$, $f'(x) = w$, $f''(x) = 0, \omega_2 = 0$, $h'(f(x)) = \frac{2}{\sigma_{\text{ID}}}(w^T x - \mu_{\text{ID}})$, $h''(f(x)) = \frac{2}{\sigma_{\text{ID}}}$. μ_{ID} and σ_{ID} denote a centroid and covariance of the representations belonging to the closest class with the target sample representation. For simplicity, we assume $\alpha = \beta = 1$ and consider a single oracle and an auxiliary sample. Under these assumptions, we can rewrite our score function as follows:

$$h(f(x_t)) - h(f(x_o)) + (h(f(x_{ta})) - h(f(x_{oa}))). \tag{C.1}$$

By utilizing Proposition 1, we can rewrite Equation C.1 as follows:

$$\begin{aligned} &h(f(x_t)) - h(f(x_o)) + h(f(x_t)) - h(f(x_o)) \\ &+ (\omega_1(x_t, x_a) - \omega_1(x_o, x_a)) \\ &+ (\omega_3(x_t, x_a) - \omega_3(x_o, x_a)). \end{aligned} \tag{C.2}$$

According to our assumption, $\omega_2 = 0$ since $f''(x) = 0$. In Equation C.2, the portion of the equation, excluding $h(f(x_t))$, represents how our methodology calibrates the score based solely on the distance-based metric:

$$\begin{aligned} &- h(f(x_o)) + h(f(x_t)) - h(f(x_o)) \\ &+ (\omega_1(x_t, x_a) - \omega_1(x_o, x_a)) \\ &+ (\omega_3(x_t, x_a) - \omega_3(x_o, x_a)) \end{aligned} \tag{C.3}$$

Let $f(x_t) - \mu_{\text{ID}} = \epsilon$, $f(x_t) - f(x_o) = d_{to}$, and $f(x_t) - f(x_a) = d_{ta}$

$$\begin{aligned} &h(f(x_t)) - 2h(f(x_o)) \\ &= h(w^T x_t) - 2h(w^T x_o) \\ &= \frac{1}{\sigma_{\text{ID}}}[(w^T x_t - \mu_{\text{ID}})^2 - 2(w^T x_o - \mu_{\text{ID}})^2] \\ &= \frac{1}{\sigma_{\text{ID}}}[\epsilon^2 - 2(\epsilon - d_{to})^2] \end{aligned} \tag{C.4}$$

$$\begin{aligned} &\omega_1(x_t, x_a) - \omega_1(x_o, x_a) \\ &= (\lambda - 1)(x_t - x_a)^T f'(x_t) h'(f(x_t)) \\ &- (\lambda - 1)(x_o - x_a)^T f'(x_o) h'(f(x_o)) \\ &= (\lambda - 1)(x_t - x_a)^T w \frac{2}{\sigma_{\text{ID}}}(w^T x_t - \mu_{\text{ID}}) \\ &- (\lambda - 1)(x_o - x_a)^T w \frac{2}{\sigma_{\text{ID}}}(w^T x_o - \mu_{\text{ID}}) \\ &= (\lambda - 1)(f(x_t) - f(x_a)) \frac{2}{\sigma_{\text{ID}}}(f(x_t) - \mu_{\text{ID}}) \\ &- (\lambda - 1)(f(x_o) - f(x_a)) \frac{2}{\sigma_{\text{ID}}}(f(x_o) - \mu_{\text{ID}}) \\ &= \frac{2(\lambda - 1)}{\sigma_{\text{ID}}}[d_{ta} \cdot \epsilon - (d_{ta} - d_{to})(\epsilon - d_{to})] \\ &= \frac{2(\lambda - 1)}{\sigma_{\text{ID}}}[d_{ta} d_{to} - d_{to}^2 + d_{to}\epsilon] \end{aligned} \tag{C.5}$$

$$\begin{aligned} &\omega_3(x_t, x_a) - \omega_3(x_o, x_a) \\ &= \frac{(\lambda - 1)^2}{2}(x_t - x_a)^T w (x_t - x_a)^T w \frac{2}{\sigma_{\text{ID}}} \\ &- \frac{(\lambda - 1)^2}{2}(x_o - x_a)^T w (x_o - x_a)^T w \frac{2}{\sigma_{\text{ID}}} \\ &= \frac{(\lambda - 1)^2}{\sigma_{\text{ID}}}[d_{ta}^2 - (d_{ta} - d_{to})^2] \\ &= \frac{(\lambda - 1)^2}{\sigma_{\text{ID}}}[2 d_{ta} d_{to} - d_{to}^2]. \end{aligned} \tag{C.6}$$

Then, we rewrite Equation C.3 as a sum of Equation C.4, C.5 and C.6:

$$\begin{aligned} &\frac{1}{\sigma_{\text{ID}}}[\epsilon^2 - 2(\epsilon - d_{to})^2] + \frac{2(\lambda - 1)}{\sigma_{\text{ID}}}[d_{ta} d_{to} - d_{to}^2 \\ &+ d_{to}\epsilon] + \frac{(\lambda - 1)^2}{\sigma_{\text{ID}}}[2 d_{ta} d_{to} - d_{to}^2]. \end{aligned} \tag{C.7}$$

The above expression must be positive when the target sample is a hard OOD sample to achieve a higher OOD score than the conventional distance-based OOD score. Equation C.4 is must be negative because $h(f(x_t)) < h(f((x_o))$:

$$\frac{1}{\sigma_{\text{ID}}}[\epsilon^2 - 2(\epsilon - d_{to})^2] < 0$$

The condition of relation between ϵ and d_{to} is

$$\epsilon < (2 - \sqrt{2}) d_{to} \text{ or } \epsilon > (2 + \sqrt{2}) d_{to}$$

Equation C.7 can be rewritten as:

$$\begin{aligned} &- \epsilon^2 + 4\epsilon d_{to} - 2 d_{to}^2 \left(1 + (\lambda - 1) + (\lambda - 1)^2\right) \\ &+ 2\epsilon d_{to}(\lambda - 1) \\ &+ 2 d_{ta} d_{to} \left((\lambda - 1) + (\lambda - 1)^2\right) > 0 \end{aligned} \tag{C.8}$$

When $\epsilon > (2+\sqrt{2})d_{to}$ and d_{to} is a positive value close to zero, there exists an auxiliary sample that satisfies Equation C.8 if:

$$d_{ta} < 0 \text{ and}$$
$$d_{ta} < \frac{(3+2\sqrt{2}+\lambda^2-6\lambda-2\sqrt{2}\lambda)d_{to}}{2\lambda(\lambda-1)},$$

□

D Validation of methodology using synthetic datasets

To complement our 1D toy experiment and further validate our theoretical analysis, we conducted experiments using 2D synthetic data generated from three Gaussian distributions. A linear model was trained on two of these distributions, forming the decision boundary, as shown in Figure 6. The third Gaussian distribution, positioned between the two learned classes, served as the OOD data. OOD data clusters locate closely to the ID data, posing a more challenging scenario.

The target OOD sample is shown in pink, the oracle sample in green, and auxiliary samples that satisfy the positive CED score condition are in orange. Consistent with our theoretical analysis, auxiliary samples that improve OOD detection tend to be positioned opposite the oracle relative to the target. Auxiliary samples located farther from the target, while still satisfying the theoretical conditions, result in higher CED scores, confirming the findings from our 1D experiment.

E Sensitivity Analysis

E.1 Hyperparameter Analysis

We validate the sensitivity of CED score to scaling parameters α and β on CLINC150 and News Category datasets for RoBERTa and LLaMA-3 8B models, as shown in Figure 7. The analysis covers $\text{FLaTS}_{\text{pre}}$+CED, KNN_{pre}+CED, and MD_{pre}+CED methods. Across all α and β values, CED consistently outperforms baseline methods for both AUROC and FPR95 metrics. This stability across parameter ranges demonstrates CED's robustness and reliability in various settings.

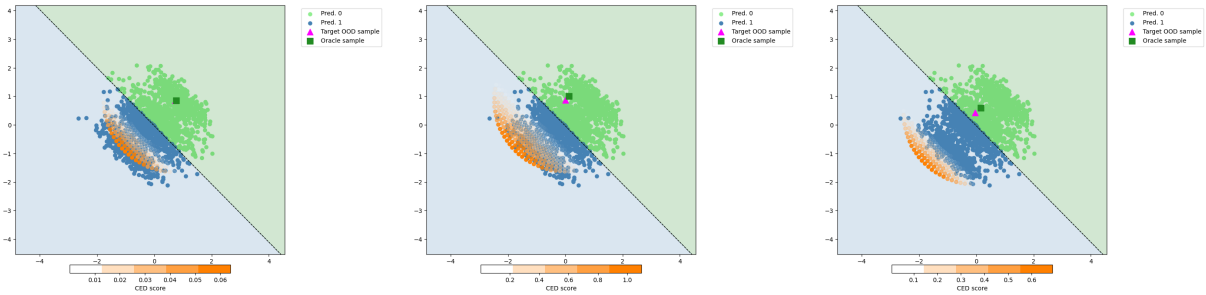
F Additional Qualitative Analysis

F.1 Case Study for OOD Detection

Additional case studies are shown in Table 5, 6. Across the Banking77 and CLINC150 datasets, we observe that combining OOD queries with auxiliary samples often leads to significant shifts in class predictions, unlike the more stable classifications seen when ID or oracle samples are combined with the same auxiliary. This pattern highlights a key distinction in how OOD and ID samples interact with auxiliary information, helping to effectively differentiate between the two.

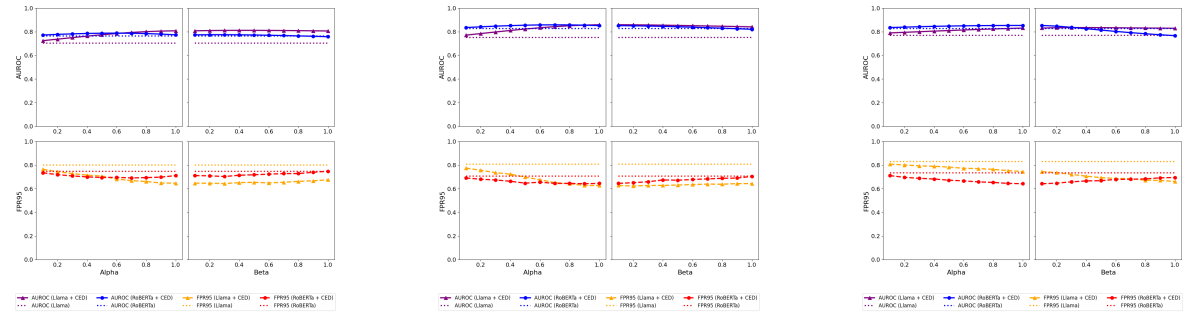
F.2 Case Study for Hallucination Detection

Each of the three boxes in Table 7 labeled Hallucination, Oracle, and Auxiliary represents the samples used to compute the CED score. In this case, CED resulted in a 15.3% increase in the OOD score compared to using the MD score as hallucination detection method. The oracle sample shares the British nationality with the hallucination sample. The auxiliary sample, introduces a broader, less relevant context. Together, these carefully chosen samples highlight contrasts that are not captured by the MD score alone, contributing to the improved OOD detection performance.

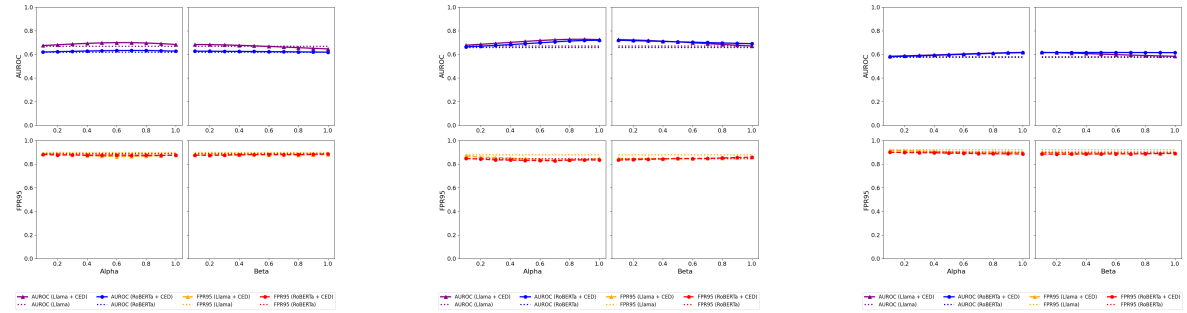


$\lambda = 0.2$ $\lambda = 0.5$ $\lambda = 0.8$

Figure 6: Theoretical analysis of CED under different λ values. The target OOD sample is shown in pink, and the oracle sample is shown in green. The results remain consistent regardless of the λ value.



(a) $\text{FLatS}_{\text{pre}}$+CED (CLINC150) (b) KNN_{pre}+CED (CLINC150) (c) MD_{pre}+CED (CLINC150)



(d) $\text{FLatS}_{\text{pre}}$+CED (News Category) (e) KNN_{pre}+CED (News Category) (f) MD_{pre}+CED (News Category)

Figure 7: Sensitivity analysis on CLINC150 and News Category datasets for RoBERTa and LLaMA-3

Sample Type	**Query Example**	**Closest Class**
OOD	I need to make a transfer, what will the fee be?	Failed Transfer
OOD + Auxiliary	I need to make a transfer, what will the fee be? I'm on vacation in Europe but I desperately need to change my PIN. Can I do this from abroad?	**Cancel Transfer**
Oracle + Auxiliary	My transfer failed, why? I'm on vacation in Europe but I desperately need to change my PIN. Can I do this from abroad?	**Failed Transfer**
ID	I ordered a card but it has not arrived. Help please!	Card arrival
ID + Auxiliary	I ordered a card but it has not arrived. Help please! I am on vacation in Spain and need to change my pin.	Change pin
Oracle + Auxiliary	I have been waiting over a week. Is the card still coming? I am on vacation in Spain and need to change my pin.	Change pin

Table 5: Examples of OOD/ID, Oracle, Auxiliary, and Combined Samples from the BANKING77 dataset.

Sample Type	Query Example	Closest Class
OOD	What are the highest-rated android phones.	Restaurant Suggestion
OOD + Auxiliary	What are the highest-rated android phones. USD to the Euro exchanges at what right now	**Transactions**
Oracle + Auxiliary	What are the best restaurants. USD to the Euro exchanges at what right now	**Restaurant Suggestion**
ID	What's the spanish word for pasta	Spelling
ID + Auxiliary	What's the spanish word for pasta. I must transfer ten dollars from my bank of america account to my capital one account.	Transfer
Oracle + Auxiliary	What's the right spelling of rambunctious. I must transfer ten dollars from my bank of america account to my capital one account.	Transfer

Table 6: Examples of OOD/ID, Oracle, Auxiliary, and Combined Samples from the CLINC150 dataset.

Sample Type	Content
Hallucination	This is a Wikipedia passage about Phillips ' Sound Recording Services. Phillips ' Sound Recording Services was a studio in the house of Percy Francis Phillips (1896 – 1984) and his family at 38 Kensington , Kensington , Liverpool , England ." 'It was the first studio in the world to be equipped with a recording console, and the first to use a microphone.
Oracle	This is a Wikipedia passage about O. G. S. Crawford. Osbert Guy Stanhope Crawford (28 October 1886 – 28 November 1957) , better known as O.' Crawford , was a British archaeologist who specialised in the study of prehistoric Britain and the archaeology of Sudan.
Auxiliary	This is a Wikipedia passage about Backmasking. Backmasking is a recording technique in which a sound or message is recorded backward onto a track that is meant to be played forward . Backmasking is a deliberate process , whereas a message found through phonetic reversal may be unintentional.

Table 7: Examples of Hallucination, Oracle, Auxiliary sample.