# Large Language Models for Anomaly and Out-of-Distribution Detection: A Survey

**Ruiyao Xu**♠ and **Kaize Ding**♠

♠Department of Statistics and Data Science, Northwestern University
ruiyaoxu2028@u.northwestern.edu, kaize.ding@northwestern.edu

## Abstract

Detecting anomalies or out-of-distribution (OOD) samples is critical for maintaining the reliability and trustworthiness of machine learning systems. Recently, Large Language Models (LLMs) have demonstrated their effectiveness not only in natural language processing but also in broader applications due to their advanced comprehension and generative capabilities. The integration of LLMs into anomaly and OOD detection marks a significant shift from the traditional paradigm in the field. This survey focuses on the problem of anomaly and OOD detection under the context of LLMs. We propose a new taxonomy to categorize existing approaches into two classes based on the role played by LLMs. Following our proposed taxonomy, we further discuss the related work under each of the categories and finally discuss potential challenges and directions for future research in this field. We also provide an up-to-date reading list [1] of relevant papers.

## 1 Introduction

Most machine learning models operate under the closed-set assumption (Krizhevsky et al., 2012), where the test data is assumed to be drawn i.i.d. from the same distribution as the training data. However, in real-world applications, this assumption often cannot hold, as test examples can come from distributions not represented in the training data. These instances, known as anomalies or out-of-distribution (OOD) samples, can severely degrade the performance and reliability of existing models (Yang et al., 2024a). To build robust AI systems, methods including probabilistic approaches (Lee et al., 2018; Leys et al., 2018) and recent deep learning techniques (Pang et al., 2021; Yang et al., 2024a) have been explored to detect these unknown instances across various domains, such as fraud de-
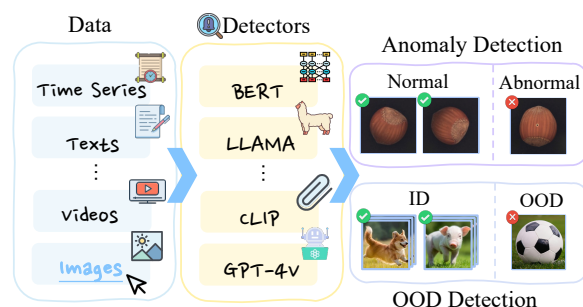


Figure 1: A simple illustration of leveraging LLMs for vision anomaly and OOD detection.

tection in finance and fault detection in industrial systems (Hilal et al., 2022; Liu et al., 2024b).

Large Language Models, such as GPT-4 (Achiam et al., 2023) and LLaMA (Touvron et al., 2023), have demonstrated remarkable capabilities in language comprehension and generation. To further harness the potential of LLMs beyond text data, there is also a growing interest in extending them to multi-modal tasks such as vision-language understanding and generation (Wang et al., 2024), evolving them into Multimodal LLMs (MLLMs) (Yin et al., 2023). Given the zero- and few-shot reasoning capabilities of LLMs and MLLMs, researchers try to apply these models to anomaly and out-of-distribution (OOD) detection, as illustrated in Figure 1, yielding promising detection results.

Remarkably, the emergence of LLMs has fundamentally changed the learning paradigm in this field. In the meantime, while leveraging LLMs to solve the problem of anomaly and OOD detection has drawn much attention, this field remains underexplored, highlighting the need for a comprehensive survey to analyze the emerging challenges and systematically review the rapidly expanding works. Recently, Salehi et al. (2021) and Yang et al. (2024a) present unified frameworks for OOD detection but do not delve into the utilization of LLMs. While Su et al. (2024) review some small-sized

---

[1] https://github.com/rux001/
Awesome-LLM-Anomaly-OOD-Detection

language models for forecasting and anomaly detection, they neither cover the usage of LLMs with emergent abilities nor discuss OOD detection. A recent survey by Miyai et al. (2024a) summarizes works on anomaly and OOD detection in vision using vision-language models but neglects other data modalities. Therefore, we aim to conduct a systematic survey that covers both anomaly and OOD detection across various data domains, concentrating on how LLMs are used in existing works.

In this survey, we propose a novel taxonomy that focuses on how LLMs can profoundly impact anomaly and OOD detection in two fundamental ways, as illustrated in Figure 2: ❶ **LLMs for Detection (§3):** We provide a detailed review of existing methods that leverage LLMs as detectors for identifying anomalies and OOD instances; and ❷**LLMs for Generation (§4):** We also review methods that utilize LLMs' emergent abilities, advanced semantic understanding, and vast knowledge to generate augmented data and explanations. At the end (**§5** and **§6**), we also summarize widely used datasets and outline future research directions, in order to provide a better understanding of anomaly and OOD detection in the era of LLMs and shed light on the following research.

## 2 Preliminaries

**Large Language Models.** Large language models (LLMs) generally refer to Transformer-based pre-trained language models with hundreds of billions of parameters or more. Early LLMs like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) utilize an encoder-only architecture, excelling in text representation learning (Bengio et al., 2013). Recently, the focus has shifted toward models aimed at natural language generation, often using the "next token prediction" objective as their core task. Examples include T5 (Raffel et al., 2020) and BART (Lewis et al., 2019), which employ an encoder-decoder structure, as well as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023), which are based on decoder-only architectures. Advancements in these architectures and training methods have led to superior reasoning and emergent abilities, such as in-context learning(Brown et al., 2020) and chain-of-thought reasoning (Wei et al., 2022).

**Multimodal Large Language Models.** The remarkable abilities of Large Language Models (LLMs) have inspired efforts to integrate language with other modalities, with a particular focus on combining language and vision. Notable examples of Multimodal Large Language Models (MLLMs) include CLIP (Radford et al., 2021), BLIP2 (Li et al., 2023a), and Flamingo (Alayrac et al., 2022), which were pre-trained on large-scale cross-modal datasets comprising images and text. Models like GPT-4(V) (OpenAI, 2023) and Gemini (Team et al., 2023) showcase the emergent abilities of Multimodal LLMs, significantly improving the performance of vision-related tasks.

### 2.1 Problem Definition

With LLMs advancing in zero-shot and few-shot learning, the general pipeline of anomaly and out-of-distribution (OOD) detection methods shifts to adapt pre-trained LLMs for detection without extensive training. This shift challenges traditional definitions of anomaly and OOD detection, as the conventional train-test paradigm may not always apply. Following previous studies (Miyai et al., 2024a; Yang et al., 2024a), we propose to redefine anomaly and OOD detection under the context of LLMs and highlight the differences between the two problems as follows:

**Definition 1** *LLM-based Anomaly Detection: Given a test dataset $\mathcal{D}_{test} = \{x_1, \cdots, x_n\}$, where each sample $x_i$ is drawn from distribution $\mathbb{P}^{in}$ or $\mathbb{P}^{out}$. The objective of LLM-based Anomaly Detection is to use a pre-trained LLM as the backbone and develop a detection model $f_{LLM}(\cdot)$ to predict whether each sample $x' \in \mathcal{D}_{test}$ belongs to $\mathbb{P}^{out}$, where $\mathbb{P}^{out}$ has covariate shift with $\mathbb{P}^{in}$*

**Definition 2** *LLM-based OOD Detection: Given a test dataset $\mathcal{D}_{test} = \{x_1, \cdots, x_n\}$, where each sample $x_i$ is drawn from distribution $\mathbb{P}^{in}$ or $\mathbb{P}^{out}$, and a known ID class set $\mathcal{C} = \{c_1, \cdots, c_k\}$. The objective of LLM-based OOD Detection is to use a pre-trained LLM as backbone and develop detection model $f_{LLM}(\cdot)$ to predict whether each sample $x' \in \mathcal{D}_{test}$ belongs to $\mathbb{P}^{out}$, where $\mathbb{P}^{out}$ has semantic shift with $\mathbb{P}^{in}$. If not, $x'$ will be classified into one of the classes in $\mathcal{C}$.*

**Discussions.** The distinction between anomaly detection and OOD detection in the context of LLMs highlights the unique challenges posed by covariate and semantic shifts. *Anomaly detection* aims to identify subtle deviations within the data that may not involve a complete change in the underlying class or concept, such as detecting defects or irregularities in industrial processes. In contrast, *OOD*
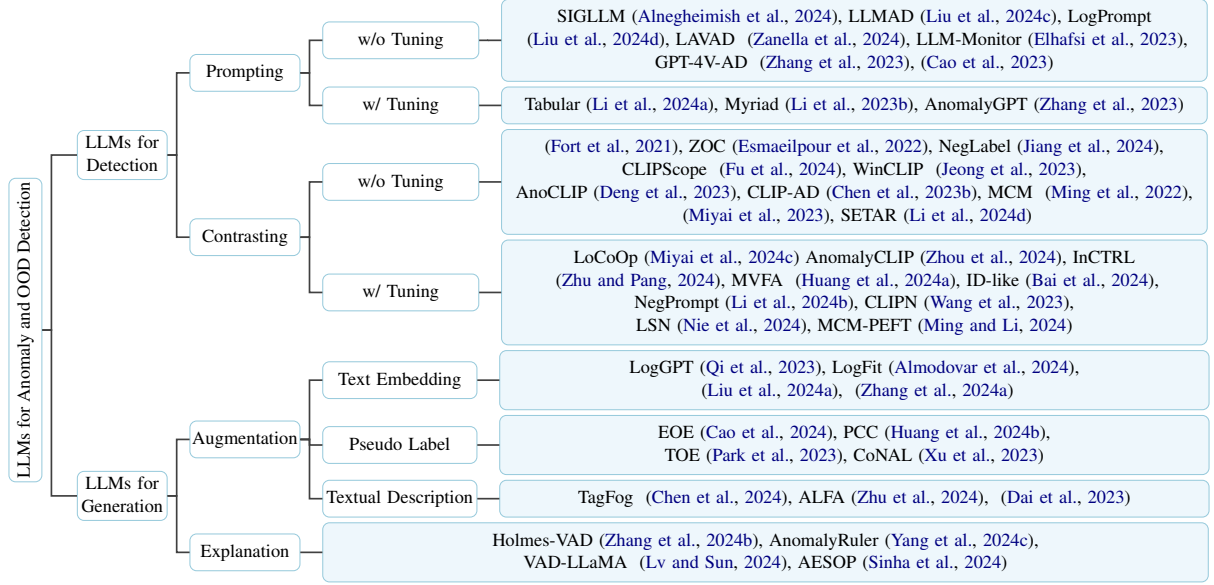
Figure 2: Taxonomy of methods utilizing LLMs for anomaly and OOD detection tasks.

*detection* focuses on identifying instances that do not belong to any of the known ID classes at the object level, such as recognizing a dog when the only provided ID class is cat. This differentiation underscores the need for tailored approaches for each detection task.

## 3 LLMs for Detection

The primary objective of this section is to explore existing works that utilize LLMs' inherent knowledge to detect anomalies or OOD samples. Under this line of research, approaches can be categorized into two classes as illustrated in Figure 3: ❶ Prompting-based Detection methods, which involve directly prompting LLMs to generate language responses that include detection results; ❷ Contrasting-based Detection methods, which focus on multimodal scenarios, using MLLMs pretrained with a contrastive objective as detectors.

### 3.1 Prompting-based Detection

The general pipeline for prompting-based detection methods consists of two primary stages: (i) constructing a structured prompt template with instruction prompt $\mathcal{P}$ and input data $\mathcal{X}$; and (ii) feeding the template-based prompt $\hat{X}$ into LLMs to generate a language response. The function $\texttt{Parse}(\cdot)$ is then applied to extract the detection results. Depending on the scenario, the LLM can either be frozen or fine-tuned, denoted as $f_{\text{LLM}}^{\heartsuit}$ or $f_{\text{LLM}}^{\clubsuit}$, respectively.

This process can be summarized as follows:

$$\text{Prompt Construction:} \quad \hat{X} = \texttt{Template}(\mathcal{X}, \mathcal{P}),$$
$$\text{Detection:} \quad \tilde{Y} = \texttt{Parse}\left(f_{\text{LLM}}^{\heartsuit/\clubsuit}(\hat{X})\right)$$

Note that the prompting-based approach primarily addresses the anomaly detection task. OOD detection research has not yet widely adopted prompting to directly identify OOD samples.

### 3.1.1 Detection without LLM Tuning

Since some approaches do not require additional tuning, they mainly focus on employing various prompt engineering techniques (Sahoo et al., 2024) to guide LLMs to produce better detection results. To design suitable prompts for anomaly detection, researchers have employed a combination of various prompt techniques, such as *role-play prompting* (Wu et al., 2023), *in-context learning* (Brown et al., 2020), and *chain-of-thought (CoT) reasoning* (Wei et al., 2022), to create effective prompt templates. Studies such as SIGLLM (Alnegheimish et al., 2024), LLMAD (Liu et al., 2024c), and LogPrompt (Liu et al., 2024d) focus on time series and log data. SIGLLM (Alnegheimish et al., 2024) investigates two distinct pipelines for using LLMs in time series anomaly detection: one directly prompts an LLM with specific role-play instructions to identify anomalous elements in given data, and the other uses the LLM's forecasting ability to detect anomalies by comparing original and forecasted signals, where discrepancies indicate

anomalies. LLMAD (Liu et al., 2024c) incorporates in-context learning examples retrieved from a constructed database and CoT prompts that inject domain knowledge of time series. LogPrompt (Liu et al., 2024d) explores three prompting strategies for log data: self-prompt, CoT prompt, and in-context prompt, demonstrating that the prompt with CoT techniques outperforms other prompting strategies. The tailored CoT prompt for log data includes a specific task instruction, i.e. "classify the given log entries into normal and abnormal categories", and step-by-step rules for considering given data as anomalies.

Unlike time series and log data which can be directly converted into raw text data, other data modalities, such as videos and images, require additional processing to be transformed into a format that LLMs can understand. For instance, LAVAD (Zanella et al., 2024) first exploits a captioning model to generate a textual description for each video frame and further uses an LLM to summarize captions within a temporal window. This summary is then used to prompt the LLM to provide an anomaly score for each frame. LLM-Monitor (Elhafsi et al., 2023) uses an object detector to identify objects in video clips and then designs specific prompt templates incorporating CoT and in-context examples to query LLMs for anomaly detection.

With the integration of multimodal understanding into LLMs, these models are now capable of comprehending various modalities beyond text, enabling more direct applications for anomaly detection across a wide range of data types. Cao et al. (2023) conduct comprehensive experiments and analyses using GPT-4V(ision) for anomaly detection across various modality datasets and tasks. To enhance GPT-4V's performance, they also incorporate different types of additional cues such as class information, human expertise, and reference images as prompts. Similarly, GPT-4V-AD (Zhang et al., 2023) employs GPT-4V as the backbone, designing a general prompt description for all image categories and injecting specific image category information, resulting in a specific output format for each region with respective anomaly scores.

### 3.1.2 Detection with LLM Tuning

Directly prompting frozen LLMs for anomaly or OOD detection results across various data types often yields suboptimal performance due to the inherent modality gap between text and other data modalities. As a result, additional training and fine-
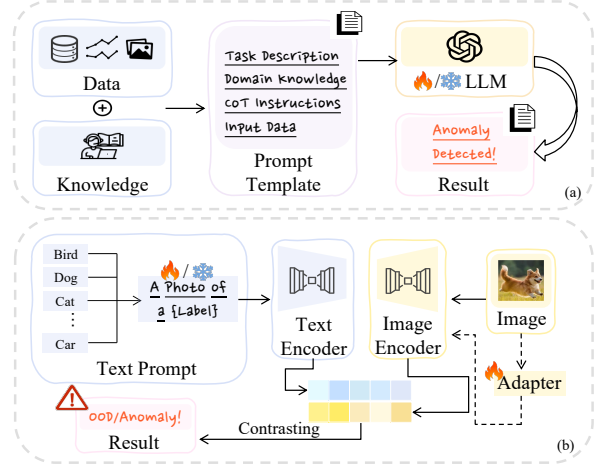


Figure 3: The illustration of two approaches in (§3): (a) Prompting-based Detection and (b) Contrasting-based Detection.

tuning on LLMs for downstream detection tasks has become a prevalent research trend. Unfortunately, fine-tuning entire LLMs is often computationally expensive and poses significant challenges. Therefore, parameter-efficient fine-tuning (PEFT) has been extensively employed instead. For example, Tabular (Li et al., 2024a) designs a prompt template to query the LLM to output anomalies based on given converted tabular data. To better adapt the LLM for anomaly detection at the batch level, they apply Low-Rank Adaptation (LoRA), using a synthetic dataset with ground truth labels in a supervised manner.

To enhance LLMs for localization understanding and adapting to industrial tasks, AnomalyGPT (Zhang et al., 2023) first derives localization features from a frozen image encoder and image decoder and these features are then fed to a tunable prompt learner. Without fine-tuning the entire LLM, they fine-tune the prompt learner with LoRA to significantly reduce computational costs. Myriad (Li et al., 2023b) employs Mini-GPT-4 as the backbone and integrates a trainable encoder, referred to as Vision Expert Tokenizer, to embed the vision expert's segmentation output into tokens that the LLM can understand. With expert-driven visual-language extraction, Myriad can generate accurate anomaly detection descriptions.

### 3.2 Contrasting-based Detection

In this section, we focus on MLLMs, such as CLIP, which are pre-trained with an image-text contrastive objective and learn by pulling the paired images and texts close and pushing others far away

in the embedding space. The zero-shot classification ability of these models further builds the foundation for contrasting-based anomaly and OOD detection methods: (i) given an image $x_i$ and a text prompt $f$ with a target class set $C$, CLIP extracts image features $h \in \mathbb{R}^D$ using an image encoder $f_{\text{img}}$, and text features $e_j \in \mathbb{R}^D$ using a text encoder $f_{\text{text}}$ with a prompt template for each class $c_j \in C$, and (ii) the similarity between $h$ and each $e_j$ is usually used as an important component in the score function $f_{\text{score}}$ for deciding whether $x_i$ is an anomaly or OOD sample. This process can be summarized as follows:

$$\text{Feature Extraction:} \quad h = f_{\text{img}}(x_i),$$
$$and \quad e_j = f_{\text{text}}(\text{prompt}(c_j)),$$
$$\text{Detection:} \quad \tilde{Y} = f_{\text{score}}\left(\cos(h, e_j)\right)$$

We further categorize contrasting-based detection methods into two classes depending on whether there exists additional training and fine-tuning.

### 3.2.1 Detection without LLM Tuning

Anomaly and OOD detection problems can indeed be understood as classification problems. Therefore, pretrained MLLMs like CLIP, with strong zero-shot classification ability, can serve as detectors themselves. By using only ID or normal prompts, CLIP can be leveraged for both OOD and anomaly detection tasks. Despite the promise, existing CLIP-like models perform zero-shot classification in a closed-world setting. That is, it will match an input into a fixed set of categories, even if it is irrelevant (Ming et al., 2022). To address this, one approach involves designing effective post-hoc score functions that rely solely on ID or normal class labels. Alternatively, some researchers incorporate anomaly or OOD class information into the text prompts, allowing the model to match OOD or abnormal images to paired prompts.

- *Without Anomaly/OOD Prompts.* For **anomaly detection**, WinCLIP (Jeong et al., 2023) initially investigates a one-class design by using only the normal prompt "normal [o]" where [o] represents object-level label, i.e "bottle", and defining an anomaly score as the similarity between vectors derived from the image encoder and normal prompts. However, this one-class design yields poorer results compared to a simple binary zero-shot framework, CLIP-AC (Jeong et al., 2023), which adapts CLIP with two class prompts: "normal [o]" vs. "anomalous [o]". These results highlight that the one-class design is less effective

for anomaly detection, and as a result, anomaly detection research generally does not follow this line with only normal prompts.

For **OOD detection**, to address the challenges of using only in-distribution (ID) class information while avoiding the matching of OOD inputs to irrelevant ID classes, one notable approach is the Maximum Concept Matching (MCM) framework proposed by (Ming et al., 2022). This method is not limited to CLIP and can be generally applicable to other pre-trained models that promote multi-modal feature alignment. They view the textual embeddings of ID classes as a collection of concept prototypes and define the maximum concept matching (MCM) score based on the cosine similarity between the image feature and the textual feature. Following the idea of MCM, several subsequent works focus on improving OOD detection results by either adding a local MCM score or modifying weights in the original MCM framework, such as (Miyai et al., 2023) and (Li et al., 2024d).

- *With Anomaly/OOD Prompts.* Fort et al. (2021) first investigate using CLIP for **OOD detection** and demonstrate encouraging performance. However, in their setup, they include the candidate labels related to the actual OOD classes and utilize this knowledge as a very weak form of outlier exposure, which contradicts the open-world assumption. Therefore, after this work, researchers aim to leverage pseudo-OOD labels in the text prompt instead of using actual OOD labels. The earliest work under this idea is ZOC (Esmaeilpour et al., 2022) which trains a text description generator on top of CLIP's image encoder to dynamically generate candidate unseen labels for each test image. The similarity of the test image with seen and generated unseen labels is used as the OOD score. Instead of training an additional text decoder, NegLabel (Jiang et al., 2024) and CLIPScope (Fu et al., 2024) rely on auxiliary datasets to gather potential OOD labels. CLIPScope gathers nouns from open-world sources as potential OOD labels and uses them in designed prompts to ensure maximal coverage of potential OOD samples. NegLabel employs the NegMining algorithm to select high-quality negative labels with sufficient semantic differences from ID labels. Recent work utilizes the emergent abilities of LLMs to generate reliable OOD labels, such as (Cao et al., 2024), (Huang et al.,

2024b), (Park et al., 2023), and (Xu et al., 2023).

For contrasting-based **anomaly detection**, following the simple binary zero-shot framework, CLIP-AC (Jeong et al., 2023), which adapts CLIP with two class prompts: "normal [o]" vs. "anomalous [o]", many subsequent research emerge. While using the default prompt has demonstrated promising performance, similar to the prompt engineering discussion around GPT-3 (Brown et al., 2020), researchers have observed that performance can be significantly improved by customizing the prompt text. Models like WinCLIP (Jeong et al., 2023) and AnoCLIP (Deng et al., 2023) use a Prompt Ensemble technique to generate all combinations of pre-defined lists of state words per label and text templates. After generating all combinations of states and templates, they compute the average of text embeddings per label to represent the normal and anomalous classes. In practice, more descriptions in prompts do not always yield better performance. Therefore, CLIP-AD (Chen et al., 2023b) proposes Representative Vector Selection (RVS), from a distributional perspective for the design of the text prompt, broadening research opportunities beyond merely crafting adjectives.

### 3.2.2 Detection with LLM Tuning

Following the similar detection pipeline of methods without LLM tuning, researchers propose to employ prompt tuning or adapter tuning techniques to eliminate the need for manually crafting prompts and enhance the understanding of local features of images. Additionally, by incorporating a few ID or normal images during training or inference phases, some methods transition into few-shot scenarios.

- *LLM Adapter-Tuning.* Adapter-tuning methods involve integrating additional components or layers into the model architecture to facilitate better alignment or localization (Hu et al., 2023). This approach is significantly useful for **anomaly detection** task, because CLIP was originally designed for classifying the semantics of objects in the scene, which does not align well with the sensory anomaly detection task where both normal and abnormal samples are often from the same class of object. To reconcile this, InCTRL (Zhu and Pang, 2024) includes a tunable adapter layer to further adapt the image representations for anomaly detection. To better adapt to medical image anomaly detection, MVFA (Huang et al.,

2024a) proposes a multi-level visual feature adaptation architecture to align CLIP's features with the requirements of anomaly detection in medical contexts. This is achieved by integrating multiple residual adapters into the pre-trained visual encoder, guided by multi-level, pixel-wise visual-language feature alignment loss functions.

- *LLM Prompt-Tuning.* Manually crafting suitable prompts always requires extensive human effort. Therefore, researchers employ the idea of prompt tuning, such as CoOp (Zhou et al., 2022), to learn a soft or differentiable context vector to replace the fixed text prompt. For **OOD detection**, most approaches rely on using auxiliary prompts to represent potential OOD textual information, and one crucial problem is to identify hard OOD data that is similar to ID samples. To solve this, Bai et al. (2024) first constructs outliers highly correlated with ID data and introduces a novel prompt learning framework for learning specific prompts for the most challenging OOD samples, which behave like ID classes. Additionally, LSN (Nie et al., 2024), NegPrompt (Li et al., 2024b), and CLIPN (Wang et al., 2023) all work on learning extra negative prompts to fully leverage the capabilities of CLIP for OOD detection. Unlike the other two approaches, CLIPN requires training an additional "no" text encoder using a large external dataset to get negative prompts for all classes. This auxiliary training is computationally expensive, limiting its application to generalized tasks. Also, LSN demonstrates that naive "no" logic prompts cannot fully leverage negative features. Therefore, both LSN and NegPrompt focus on training on more detailed negative prompts, while LSN also aims to develop class-specific positive and negative prompts, enabling more accurate detection.

Instead of focusing on leveraging OOD information into the text encoder, some methods aim to perform prompt tuning to optimize word embeddings for ID labels and then use the MCM score as the detection criterion. MCM-PEFT (Ming and Li, 2024) demonstrates that simply applying prompt tuning for CLIP on few-shot ID datasets can significantly improve detection accuracy. However, a primary limitation of this approach is its exclusive reliance on the features of ID classes, leading to incorrect detection when input images share a high visual similarity with the class in the prompt. To address this, LoCoOp

(Miyai et al., 2024c) treats such ID-irrelevant nuisances as OOD and learns to push them away from the ID class text embeddings, preventing the model from producing high ID confidence scores for the OOD features. Additionally, Lafon et al. (2024) enhances detection capabilities by learning a diverse set of prompts utilizing both global and local visual representations. For **anomaly detection** which emphasizes more on learning local features, AnomalyCLIP (Zhou et al., 2024) aims to learn object-agnostic text prompts that capture generic normality and abnormality in images, allowing the model to focus on abnormal regions rather than object semantics.

## 4 LLMs for Generation

In this section, we review methods that leverage LLMs as generative tools for enhancing anomaly and OOD detection. LLMs use their extensive pre-trained knowledge to generate augmented data, such as embeddings, pseudo labels, and textual descriptions, improving detection performance (Li et al., 2024c; Ding et al., 2024). Furthermore, due to their ability to understand and generate human-like text, LLMs have been explored for providing insightful explanations and analyses of detection results, aiding in interpretation, planning, and decision-making. These methods are classified into two main approaches: ❶ Augmentation-centric Generation and ❷ Explanation-centric Generation.

### 4.1 Augmentation-centric Generation

LLMs serve as effective tools for data augmentation in anomaly and OOD detection tasks by generating textual embeddings, pseudo labels, and descriptive text. The extensive pre-training of LLMs on large datasets and the autoregressive training objective endow them with superior generative capabilities. These capabilities allow LLMs to produce high-quality embeddings, create synthetic labels, and provide additional descriptive information, ultimately boosting the performance and robustness of detection models.

### 4.1.1 Text Embedding-based Augmentation

LLMs are highly effective feature extractors, producing meaningful embeddings that can be used in detection tasks. This augmentation enables detection models to capture more subtle patterns and distinctions, leading to more accurate and robust detection performance. For instance, Hadadi et al. (2024) and Qi et al. (2023) fine-tune pre-trained

GPT models on log data and use the extracted semantic embeddings as key components for future anomaly detection.

For OOD detection in text data, a common approach involves using encoder-only LLMs to generate sentence representations that are used to compute OOD confidence scores (Liu et al., 2024a). These models are typically fine-tuned on ID data, and OOD detectors are applied to the generated representations. Recently, there has been a shift toward leveraging larger language models with decoder architectures, which provide enhanced capabilities in refining textual representations. Liu et al. (2024a) explore the use of decoder-only LLMs, such as LLaMa, incorporating fine-tuning techniques like LoRA to reduce the additional parameters. Their findings demonstrate that fine-tuned LLMs, combined with customized OOD scoring functions, significantly improve OOD detection performance. A key advantage of decoder-based LLMs is their autoregressive ability, which allows for more effective handling of sequential data. Building on this, Zhang et al. (2024a) propose using the likelihood ratio between a pre-trained and fine-tuned LLM as a criterion for OOD detection, leveraging the deep contextual understanding embedded within LLMs for text data.

### 4.1.2 Pseudo Label-based Augmentation

The emergent capabilities of LLMs provide a promising approach for generating high-quality synthetic datasets, including pseudo labels for OOD samples. A significant challenge in OOD detection is the lack of labeled OOD data, which can limit model performance. Traditionally, obtaining OOD labels requires extensive human effort, but LLMs can mitigate this by generating pseudo-OOD labels through carefully designed prompts.

For example, EOE (Cao et al., 2024) and PCC (Huang et al., 2024b) prompt LLMs to generate visually similar OOD class labels, which are then used to define a new scoring function. This approach significantly outperforms methods relying solely on known ID labels. TOE (Park et al., 2023) further evaluates the generation of pseudo-OOD labels at three verbosity levels—word-level, description-level, and caption-level—using BERT, GPT-3, and BLIP-2, respectively. Results indicate that caption-level pseudo-OOD labels generated by BLIP-2, which incorporates both semantic and visual understanding, perform the best. In text data, CoNAL (Xu et al., 2023) prompts LLMs to
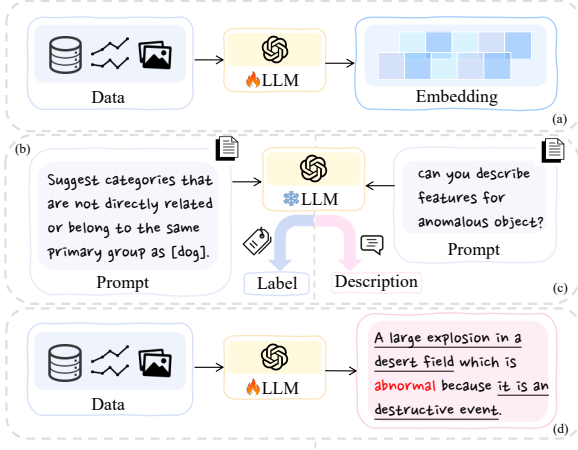
Figure 4: The illustration of four approaches in (§4): (a) Text Embedding-based Augmentation; (b) Pseudo Label-based Augmentation; (c) Textual Description-based Augmentation; and (d) Explanation-centric Generation.

extend closed-set labels with novel examples and generates a comprehensive set of probable OOD samples. By applying contrastive confidence loss during training, the model achieves high accuracy on the ID training set while maintaining lower confidence on the generated OOD examples.

### 4.1.3 Textual Description-based Augmentation

In addition to generating pseudo labels, LLMs are also used to generate textual descriptions of both known ID classes and potential OOD samples. For example, TagFog (Chen et al., 2024) employs a Jigsaw strategy to generate fake OOD samples and prompts ChatGPT to create detailed descriptions for each ID class, guiding the training of the image encoder in CLIP for OOD detection. In anomaly detection tasks, it is essential for LLMs to recognize the close correlation between normal images and their respective prompts, while identifying a more distant association with abnormal prompts. This requires detailed and nuanced descriptions of normal and anomalous stages of objects. ALFA (Zhu et al., 2024) formulates prompts for LLMs to describe both normal and abnormal features for each class, and these descriptions are then used to improve the detection of abnormal objects. To avoid LLM hallucination, Dai et al. (2023) introduce a consistency-based uncertainty calibration method, where LLMs describe visual features for distinguishing categories in images, and the confidence score of each generation is estimated accordingly.

### 4.2 Explanation-centric Generation

Beyond augmentation, LLMs' powerful reasoning and natural language generation abilities allow them to provide insightful explanations for anomaly and OOD detection outcomes. These explanations are especially important in safety-critical domains, such as autonomous driving, where transparency and interpretability are crucial.

For example, Holmes-VAD (Zhang et al., 2024b) trains a lightweight temporal sampler to select frames with high anomaly scores and uses an LLM to generate detailed explanations, providing clear insights into the detected anomalies. VAD-LLaMA (Lv and Sun, 2024) generates instruction-tuning data to train the projection layer of Video-LLaMA, enabling more comprehensive explanations of anomalies. AnomalyRuler (Yang et al., 2024c) employs a rule-based reasoning strategy with few-normal-shot prompting, providing interpretable, rule-driven explanations that can quickly adapt to various video anomaly detection scenarios.

Additionally, LLMs are being used in autonomous agents to guide decision-making after anomaly or OOD detection. For instance, AESOP (Sinha et al., 2024) leverages the autoregressive capabilities of an LLM to provide zero-shot assessments on whether interventions are required in robotic systems after an anomaly is detected. By utilizing LLMs' generative reasoning, these systems can plan and respond to anomalies in an efficient and informed manner.

## 5 Evaluation Datasets

In this section, we introduce commonly used datasets across multiple modalities in Table 1, including images, videos, text, and time series, that serve as benchmarks for anomaly and OOD detection We also introduce widely used evaluation metrics in Appendix E.

**Images.** Industrial anomaly detection is commonly evaluated using MVTec-AD (Bergmann et al., 2019) and VisA (Zou et al., 2022), which provide benchmarks for defect detection and localization. In the medical domain, datasets such as Chest X-ray (Kermany et al.) and Head CT (Felipe, 2018) are used for detecting abnormalities in clinical imaging. Logical anomaly detection is assessed with MVTec LOCO (Bergmann et al., 2022). For OOD detection, which focuses on semantic shifts, large-scale datasets such as ImageNet (Deng et al., 2009), iNaturalist (Van Horn et al., 2018),

| | Task Type | Datasets |
|---|---|---|
| **Images** | Anomaly | MVTec-AD (Bergmann et al., 2019) |
| | Anomaly | VisA (Zou et al., 2022) |
| | Anomaly | Chest X-ray (Kermany et al.) |
| | Anomaly | Head CT (Felipe, 2018) |
| | Anomaly | LOCO (Bergmann et al., 2022) |
| | OOD | ImageNet (Deng et al., 2009) |
| | OOD | iNaturalist (Van Horn et al., 2018) |
| | OOD | Places (Zhou et al., 2017) |
| | OOD | SUN (Xiao et al., 2010) |
| | OOD | Texture (Cimpoi et al., 2014) |
| **Videos** | Anomaly | UCF-Crime (Sultani et al., 2018) |
| | Anomaly | XD-Violence (Wu et al.) |
| | Anomaly | VAD-Intruct50k (Zhang et al., 2024b) |
| **Texts** | Anomaly | NLP-AD (Li et al., 2024e) |
| | Anomaly | AD-NLP (Bejan et al., 2023) |
| | OOD | 20NG (Lang, 1995) |
| | OOD | SST-2 (Socher et al., 2013) |
| | OOD | CLINC150 (Repository, 2020) |
| **Times Series** | Anomaly | ODDS (Rayana, 2016) |
| | Anomaly | Yahoo (Paparrizos et al., 2022) |
| | Anomaly | ECG (Paparrizos et al., 2022) |
| | Anomaly | SVDB (Paparrizos et al., 2022) |
| | Anomaly | IOPS (Paparrizos et al., 2022) |

Table 1: A summary of datasets for Anomaly and Out-of-Distribution (OOD) Detection by Modality.

Places (Zhou et al., 2017), and SUN (Xiao et al., 2010) are widely used.

**Videos.** UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al.) are widely used for detecting anomalous activities in surveillance footage. The VAD-Instruct50k dataset (Zhang et al., 2024b) introduces a large-scale multimodal benchmark for video anomaly detection.

**Texts.** Textual anomaly and OOD detection have gained attention with the rise of LLMs. NLP-AD (Li et al., 2024e) and AD-NLP (Bejan et al., 2023) provide benchmarks for detecting anomalous patterns in textual data. Recent work, AD-LLM (Yang et al., 2024b), evaluates LLM performance on text anomaly detection, demonstrating their superior effectiveness in this task. Textual OOD detection is commonly assessed using CLINC150 (Repository, 2020), which partitions intent categories into ID and OOD classes. Additional benchmarks, such as 20 Newsgroups (20NG) (Lang, 1995) and SST-2 (Socher et al., 2013), are used for OOD detection in topic classification and sentiment analysis.

**Time Series.** The ODDS dataset (Rayana, 2016) serves as a standard benchmark for time series anomaly deteciton, while Yahoo, ECG, SVDB, and IOPS datasets (Paparrizos et al., 2022) focus on detecting anomalies in financial transactions, phys-iological signals, and system performance.

## 6 Challenges and Future Directions

In this section, we briefly summarize challenges and future directions within the anomaly and OOD detection research field in the era of LLMs.

**Explainability and Trustworthiness.** In addition to accurately detecting anomalies or OOD samples, there is an increasing trend to utilize LLMs to provide reasonable explanations and serve as agents to plan future actions. Future research should focus on developing methods to enhance the explainability of LLMs for anomaly or OOD detection, increasing the trustworthiness of LLM-based systems and facilitating their adoption in critical domains such as healthcare and finance (Holzinger et al., 2019; Guidotti et al., 2019).

**Unsolvable Problem Detection.** Miyai et al. (2024b) propose Unsolvable Problem Detection (UPD), which evaluates the LLMs' ability to recognize and abstain from answering unexpected or unsolvable input questions, aiding in preventing incorrect or misleading outputs in critical applications where the consequences of errors can be significant. Future work should focus on incorporating the concepts of OOD detection techniques for solving the UPD problems.

**Handling Multimodal Data.** The emergence of MLLMs capable of processing and understanding multiple data modalities offers significant potential in the field of anomaly and OOD detection (Alayrac et al., 2022; Li et al., 2023a). Future research should explore methods to better adapt LLMs to comprehend and integrate various multimodal data, thereby enhancing their ability to detect anomalies and OOD instances across diverse datasets.

## 7 Conclusion

In this survey, we examine the use of Large Language Models (LLMs) and multimodal LLMs (MLLMs) in anomaly and out-of-distribution (OOD) detection. We introduce a novel taxonomy categorizing methods into two approaches based on the role of LLMs in the architectures: detection, and generation. This taxonomy clarifies how LLMs can augment data, detect anomalies or OOD, and build explainable systems. We also summarize commonly used datasets and discuss future research directions, aiming to highlight advancements and challenges in the field of anomaly or OOD detection and encourage further progress.

## Limitations

While this survey provides a comprehensive overview of the utilization of Large Language Models (LLMs) for anomaly and out-of-distribution (OOD) detection, several limitations should be acknowledged:

- **Scope of Coverage**: Although we endeavored to include the latest research, the rapid pace of advancements in the field means that some recent developments may not be covered.

- **Depth of Analysis**: Given the broad range of topics discussed, certain methods may not be explored in the depth they deserve.

- **Evaluations and Benchmarks**: Due to space constraints, we did not include a detailed summary of common evaluation metrics and benchmark datasets used in this area.

By acknowledging these limitations, we aim to provide a balanced perspective and encourage further research to address these gaps and build on the foundations laid by this survey.

## References

Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. 2019. Latent space autoregression for novelty detection. In *CVPR*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*.

Crispin Almodovar, Fariza Sabrina, Sarvnaz Karimi, and Salahuddin Azad. 2024. Logfit: Log anomaly detection using fine-tuned language models. *IEEE Transactions on Network and Service Management*.

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

Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. 2024. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*.

Matei Bejan, Andrei Manolache, and Marius Popescu. 2023. AD-NLP: A benchmark for anomaly detection in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10766–10778, Singapore. Association for Computational Linguistics.

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. 2022. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*.

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. *NeurIPS*.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. 2024. Envisioning outlier exposure by large language models for out-of-distribution detection. In *ICML*.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. 2023. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*.

Jiankang Chen, Tong Zhang, Wei-Shi Zheng, and Ruixuan Wang. 2024. Tagfog: Textual anchor guidance and fake outlier generation for visual out-of-distribution detection. In *AAAI*.

Xuhai Chen, Yue Han, and Jiangning Zhang. 2023a. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*.

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Yong Liu. 2023b. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *arXiv preprint arXiv:2311.00453*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *JMLR*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *CVPR*.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. 2024. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. *arXiv preprint arXiv:2405.14529*.

Gaudenz Danuser and Markus Stricker. 1998. Parametric model fitting: From inlier characterization to outlier detection. *TPAMI*.

Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. 2018. Image anomaly detection with generative adversarial networks. In *ECML&KDD*.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer.

Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. 2023. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa AD Nesnas, and Marco Pavone. 2023. Semantic anomaly detection with large language models. *Autonomous Robots*.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI*.

Kitamura Felipe. 2018. Head ct - hemorrhage. https://www.kaggle.com/datasets/felipekitamura/headct-hemorrhage.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. 2021. Exploring the limits of out-of-distribution detection. In *NeurIPS*.

Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. 2024. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. *arXiv preprint arXiv:2405.14737*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. 2024a. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. *arXiv preprint arXiv:2404.13671*.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024b. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*.

Fatemeh Hadadi, Qinghua Xu, Domenico Bianculli, and Lionel Briand. 2024. Anomaly detection on unstable logs with gpt models. *arXiv preprint arXiv:2406.07467*.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Waleed Hilal, S. Andrew Gadsden, and John Yawney. 2022. Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*.

Andreas Holzinger, Georg Langs, Daniel Denk, Kurt Zatloukal, and Henning Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Weiming Hu, Jun Gao, Bing Li, Ou Wu, Junping Du, and Stephen Maybank. 2018. Anomaly detection using local kernel density estimation and context-based regression. *TKDE*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. 2024a. Adapting visual-language models for generalizable anomaly detection in medical images. In *CVPR*.

K Huang, G Song, Hanwen Su, and Jiyan Wang. 2024b. Out-of-distribution detection using peer-class generated by large language model. *arXiv preprint arXiv:2403.13324*.

Kristen Jaskie and Andreas Spanias. 2019. Positive and unlabeled learning algorithms and applications: A survey. In *International Conference on Information, Intelligence, Systems and Applications*.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2024. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*.

Diederik P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *NeurIPS*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. 2020. Normalizing flows: An introduction and review of current methods. *TPAMI*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*.

Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. 2024. Gallop: Learning global and local prompts for vision-language models. *arXiv preprint arXiv:2407.01400*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*. Morgan Kaufmann, San Francisco (CA).

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Christophe Leys, Olivier Klein, Yves Dominicy, and Christophe Ley. 2018. Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of Experimental Social Psychology*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Li Fei-Fei. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. 2024b. Learning transferable negative prompts for out-of-distribution detection. In *CVPR*.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024c. Empowering large language models for textual data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12734–12751, Bangkok, Thailand. Association for Computational Linguistics.

Yixia Li, Boya Xiong, Guanhua Chen, and Yun Chen. 2024d. Setar: Out-of-distribution detection with selective low-rank approximation. *arXiv preprint arXiv:2406.12629*.

Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, and Yue Zhao. 2024e. Nlpadbench: Nlp anomaly detection benchmark. *arXiv preprint arXiv:2412.04784*.

Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. 2023b. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

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.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. 2024b. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *NeurIPS*.

Yilun Liu, Shimin Tao, Weibin Meng, Feiyu Yao, Xiaofeng Zhao, and Hao Yang. 2024d. Logprompt: Prompt engineering towards zero-shot and interpretable log analysis. In *ICSE*.

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*.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. 2022. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *arXiv preprint arXiv:2205.11474*.

Hui Lv and Qianru Sun. 2024. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. 2022. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*.

Yifei Ming and Yixuan Li. 2024. How does fine-tuning impact out-of-distribution detection for vision-language models? *IJCV*.

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

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. 2024b. Unsolvable problem detection: Evaluating trustworthiness of vision language models. *arXiv preprint arXiv:2403.20331*.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. 2023. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. 2024c. Locoop: Few-shot out-of-distribution detection via prompt learning. *NeurIPS*.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. 2024. Out-of-distribution detection with negative prompts. In *ICLR*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proc. VLDB Endow.*

Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. 2023. On the powerfulness of textual outlier exposure for visual ood detection. In *NeurIPS*.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*.

Stanislav Pidhorskyi, Ranya Almohsen, Donald A Adjeroh, and Gianfranco Doretto. 2018. Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*.

Jiaxing Qi, Shaohan Huang, Zhongzhi Luan, Shu Yang, Carol Fung, Hailong Yang, Depei Qian, Jing Shang, Zhiwen Xiao, and Zhihui Wu. 2023. Loggpt: Exploring chatgpt for log-based anomaly detection. In *HPCC/DSS/SmartCity/DependSys*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Shebuti Rayana. 2016. Odds library.

UCI Machine Learning Repository. 2020. CLINC150.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards total recall in industrial anomaly detection. In *CVPR*.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *ICML*.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. 2018. Adversarially learned one-class classifier for novelty detection. In *CVPR*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. 2021. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*.

Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. 2024. Real-time anomaly detection and re-active planning with large language models. *arXiv preprint arXiv:2407.08735*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *CVPR*.

Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. In *NeurIPS*.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *ICML*.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*.

David Martinus Johannes Tax. 2002. One-class classification: Concept learning in the absence of counter-examples.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jing Tian, Michael H Azarian, and Michael Pecht. 2014. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *PHM Society European Conference*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Thomas Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Ferhan Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Melissa Turcotte, Juston Moore, Nick Heard, and Aaron McPhall. 2016. Poisson factorization for peer-based anomaly detection. In *IEEE Conference on Intelligence and Security Informatics (ISI)*.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *CVPR*.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. In *ICML*.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In *NLPCC*.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Albert Xu, Xiang Ren, and Robin Jia. 2023. Contrastive novelty-augmented learning: Anticipating outliers with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11778–11801, Toronto, Canada. Association for Computational Linguistics.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024a. Generalized out-of-distribution detection: A survey. *IJCV*.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, et al. 2024b. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. 2024c. Follow the rules: Reasoning for video anomaly detection with large language models. *arXiv preprint arXiv:2407.10299*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *CVPR*.

Andi Zhang, Tim Z Xiao, Weiyang Liu, Robert Bamler, and Damon Wischik. 2024a. Your finetuned large language model is already a powerful out-of-distribution detector. *arXiv preprint arXiv:2404.08679*.

Bangzuo Zhang and Wanli Zuo. 2008. Learning from positive and unlabeled examples: A survey. In *International Symposiums on Information Processing*.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024b. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*.

Jiangning Zhang, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, and Yong Liu. 2023. Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. *arXiv preprint arXiv:2311.02612*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *TPAMI*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *IJCV*.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2024. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*.

Yibo Zhou. 2022. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *CVPR*.

Jiaqi Zhu, Shaofeng Cai, Fang Deng, and Junran Wu. 2024. Do llms understand visual anomalies? uncovering llm capabilities in zero-shot anomaly detection. In *ACM Multimedia 2024*.

Jiawen Zhu and Guansong Pang. 2024. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*.

Ev Zisselman and Aviv Tamar. 2020. Deep residual flow for out of distribution detection. In *CVPR*.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*.

## A  LLM-based Anomaly and OOD Detection: Strengths and Challenges

### A.1  Strengths

**Zero-shot/Few-shot:** Traditional anomaly and OOD detection methods require extensive training on well-defined normal and in-distribution datasets, which can be both time-consuming and computationally expensive. In contrast, LLMs can perform zero-shot or few-shot reasoning or learning, producing detection results without needing large-scale training (Kojima et al., 2022). This allows researchers to bypass the complex process of data collection and model training, enabling faster anomaly and OOD detection.

**Explainability and Interpretability:** LLMs possess strong reasoning abilities that can contribute to building explainable systems for anomaly and OOD detection. Traditional methods often rely on scores that offer little insight into the detection process. In contrast, LLMs can provide detailed, interpretable explanations for their detection results, offering valuable insights for future actions. Furthermore, LLMs can be integrated as agents within a system, assisting in planning the next steps when an anomaly or OOD event is detected.

### A.2  Challenges

**Computational Efficiency and Token Limits:** A major concern when leveraging LLMs for anomaly and OOD detection is computational inefficiency. Applying LLMs in these tasks often requires complex reasoning which can lead to significant computational overhead. Additionally, many LLMs have input token limits, making it impossible to feed large amounts of data directly into the model. To address this, researchers must carefully design architectures that allow LLMs to process the data effectively. For instance, techniques like Retrieval-Augmented Generation have been used to retrieve the most relevant data to avoid token limit issues (Liu et al., 2024c). Moreover, methods such as model pruning and knowledge distillation should be considered to reduce computational costs while maintaining high accuracy.

**Domain Knowledge:** While LLMs are trained on vast and diverse datasets, they may lack specific domain expertise needed for certain anomaly and OOD detection tasks. To enhance performance in these specialized domains, incorporating domain knowledge into the LLMs is crucial. One strategy is injecting domain knowledge into prompts to guide the LLM's understanding. Another approach involves using adapters or fine-tuning to better tailor the LLMs to domain-specific problems, ensuring they perform well in specialized tasks.

**Hallucination and Trustworthiness:** LLMs can sometimes produce inaccurate or fabricated information, a phenomenon known as hallucination. In the context of anomaly and OOD detection, hallucinations pose a significant risk, potentially leading to incorrect or misleading results. To mitigate this, researchers need to work on reducing hallucination rates and improving the trustworthiness of the model. Manual checks may still be necessary in critical applications, as LLMs should be seen as assistants rather than sole decision-makers.

## B  Anomaly Detection Research Roadmap

Anomaly detection has evolved significantly from traditional statistical methods to deep learning approaches and more recently to Large Language Model (LLM)-based methods. These advancements have expanded the range of applications across diverse data modalities and downstream tasks. Following Yang et al. (2024a), traditional methodologies can be grouped into density-based, reconstruction-based, distance-based, and classification-based methods. Below is an overview of these traditional approaches, followed by a discussion of the advantages and challenges of LLM-based anomaly detection.

### B.1  Traditional Methods

**Density-based Methods.** Density-based methods model the distribution of normal data and detect anomalies by evaluating how well a sample fits this modeled distribution. The underlying assumption is that normal data is more likely to have a higher likelihood under the distribution, while anomalous data will have a lower likelihood.

Parametric density estimation assumes a predefined form for the distribution, such as a multivariate Gaussian or Poisson distribution (Danuser and Stricker, 1998; Leys et al., 2018; Turcotte et al., 2016). These methods perform well when the data distribution adheres to the parametric assumption but can struggle with more complex cases. Nonparametric density estimation methods, such as histograms and kernel density estimation (KDE), offer more flexibility by not assuming a fixed parametric form, making them suitable for handling more complex distributions (Parzen, 1962; Hu et al.,

2018). Modern deep learning techniques enhance density estimation by learning high-quality feature representations. Methods such as autoencoders (AE), variational autoencoders (VAE), and flow-based models are commonly used (Kramer, 1991; Kingma and Welling, 2013; Goodfellow et al., 2014).

**Reconstruction-based Methods.** Reconstruction-based methods operate on the assumption that models trained on normal data will reconstruct those samples accurately, while anomalous data will result in higher reconstruction errors. This discrepancy in reconstruction performance is used to identify anomalies. Common approaches include sparse reconstruction, where normal samples are represented by a small set of basis functions, while anomalies are not. Autoencoders and variational autoencoders are often employed to capture these differences in reconstruction errors (Kramer, 1991; Kingma and Welling, 2013).

Recent advancements have sought to reduce the computational costs associated with reconstruction-based methods, for example by focusing on reconstructing hidden features or masking parts of the input (Pidhorskyi et al., 2018). These improvements enhance both the accuracy and efficiency of anomaly detection, without requiring pixel-level reconstruction.

**Distance-based Methods.** Distance-based methods detect anomalies by measuring the distance between test samples and reference points, such as class prototypes or centroids. Anomalies are expected to be further from these reference points than normal samples (Tian et al., 2014).

**Classification-based Methods.** Classification-based methods treat anomaly detection as a supervised learning problem. In one-class classification (OCC) (Tax, 2002), the goal is to learn a decision boundary that encompasses the normal data, with any data points outside this boundary being classified as anomalies.

DeepSVDD (Ruff et al., 2018) is a notable deep learning method for OCC, where a deep network learns a compact representation of the normal class. Semi-supervised approaches, such as positive-unlabeled (PU) learning (Zhang and Zuo, 2008; Bekker and Davis, 2020; Jaskie and Spanias, 2019), are also used when only a subset of the normal data is labeled, with the rest being unlabeled. Self-supervised learning methods have been proposed as well, using pretext tasks such as contrastive learning or future frame prediction to iden-

tify anomalies (Tack et al., 2020).

## B.2 LLM-based Anomaly Detection

The advent of Large Language Models (LLMs) like GPT, as well as multimodal LLMs, has introduced new possibilities for anomaly detection. LLMs offer zero-shot and few-shot learning capabilities, allowing anomalies to be detected with minimal or no task-specific training. This is particularly valuable in scenarios where labeled data is scarce or unavailable. LLM-based approaches benefit from their pre-trained knowledge and can adapt to various data modalities, including images, and videos. By using natural language processing capabilities, LLMs can provide explainability in anomaly detection, offering reasoning as to why a particular instance is flagged as anomalous.

However, LLM-based methods come with certain challenges, including high computational costs and limitations related to token size. These models are computationally intensive to run and may struggle with long input sequences, which necessitates techniques such as retrieval-augmented generation (RAG) or model pruning to manage these constraints.

## B.3 Comparison: Traditional vs. LLM-based Anomaly Detection

When comparing traditional and LLM-based anomaly detection methods, several key differences emerge:

- *Assumptions*: Traditional methods often rely on predefined assumptions about the data distribution, such as the parametric forms used in density-based methods. In contrast, LLM-based approaches are better equipped to generalize across a variety of tasks.

- *Data Requirements*: Traditional methods, particularly those based on deep learning, usually require large labeled datasets for training. LLM-based methods excel in zero-shot or few-shot settings, enabling them to detect anomalies with minimal task-specific data.

- *Explainability*: Traditional methods lack the ability to explain their decisions in natural language. LLM-based approaches can not only detect anomalies but also provide natural language explanations, which improves transparency and trustworthiness for further steps.

- *Computational Efficiency*: Traditional methods are generally more computationally efficient

compared to LLM-based methods, especially when fine-tuning models. However, LLM-based approaches offer greater flexibility and can handle a wider range of tasks, though at the cost of higher computational resources. Moreover, LLMs do not require extensive data preparation and training from scratch, which can offset the computational overhead in certain scenarios.

- *Generalization*: LLM-based methods are highly adaptable, capable of processing different types of data, such as text and images. In contrast, traditional methods often need custom architectures tailored to the specific data modality.

In conclusion, while traditional anomaly detection methods remain effective and computationally efficient, LLM-based methods provide greater flexibility, generalization, and explainability. This makes LLM-based approaches increasingly valuable in modern, complex anomaly detection tasks.

## C  Out-of-Distribution Detection Research Roadmap

Compared to Anomaly Detection (AD), Out-of-distribution (OOD) detection emerged in 2017 and has since received increasing attention. OOD detection is critical for ensuring the reliability and safety of machine learning models by identifying samples that fall outside the distribution of the training data. Following Yang et al. (2024a), traditional OOD detection methods can be categorized into classification-based, density-based, distance-based, and reconstruction-based methods. These approaches vary in how they define and detect OOD samples, with each showing strengths depending on data characteristics and the task at hand.

### C.1  Traditional Methods

**Classification-based Methods.** Classification-based OOD detection methods rely on the outputs of neural networks , typically using the softmax probabilities of a classifier to determine whether a sample is in-distribution (ID) or OOD. The most common baseline is the Maximum Softmax Probability (MSP) method, which flags samples with lower softmax scores as OOD (Hendrycks and Gimpel, 2017). This has led to more advanced techniques that either post-process the classification outputs or modify the training process to improve OOD detection performance (Liu et al., 2020). Given its alignment with classification tasks, this

approach remains one of the most prominent methods for OOD detection.

**Density-based Methods.** Density-based methods explicitly model the distribution of in-distribution (ID) data using probabilistic models, assuming that OOD samples will lie in low-density regions (Zong et al., 2018; Abati et al., 2019; Pidhorskyi et al., 2018; Deecke et al., 2018; Sabokrou et al., 2018). Techniques such as class-conditional Gaussian models allow for probabilistic modeling of ID classes, while flow-based models are also used for density estimation (Kobyzev et al., 2020; Zisselman and Tamar, 2020; Kingma and Dhariwal, 2018; Van Oord et al., 2016). However, density-based methods sometimes assign higher likelihoods to OOD samples, leading to challenges in reliability. Methods such as likelihood ratio-based approaches and ensembles have been proposed to address these issues, though these approaches tend to be computationally expensive and often fall behind classification-based methods in performance.

**Distance-based Methods.** Distance-based methods operate under the assumption that OOD samples are farther from the centroids or prototypes of in-distribution classes in feature space. A popular parametric approach is to use Mahalanobis distance to compute the distance between test samples and class centroids (Lee et al., 2018), whereas non-parametric methods are increasingly favored for their flexibility and simplicity (Sun et al., 2022). These methods use various distance metrics—such as Euclidean distance and geodesic distance—to detect OOD samples.

**Reconstruction-based Methods.** Reconstruction-based methods leverage encoder-decoder models to reconstruct input samples and detect OOD samples by measuring reconstruction error. The premise is that models trained on ID data will exhibit lower reconstruction errors for ID samples and higher errors for OOD samples. (Zhou, 2022).

### C.2  LLM-based OOD Detection

Large Language Models (LLMs) and Multi-Modal LLMs (MLLMs) have transformed Out-of-Distribution (OOD) detection by leveraging pre-trained models like CLIP to perform downstream detection tasks. These models are capable of detecting OOD samples in zero-shot or few-shot settings, meaning they can generalize to unseen data with little to no additional training. This represents a shift from traditional OOD detection methods, which typically rely on training classifiers using the entire

in-distribution (ID) dataset.

Incorporating the internal knowledge of pre-trained MLLMs, the field is progressing towards even greater computational efficiency, where minimal or no training data is needed. This ability to operate with limited data while maintaining performance makes LLM-based OOD detection especially appealing for real-world applications.

## C.3 Comparison of Traditional vs. LLM-based OOD Detection

The shift from traditional OOD detection methods to LLM-based approaches marks a fundamental change in how OOD detection is defined and executed. Traditional methods, such as classification-based approaches like Maximum Softmax Probability (MSP) or distance-based techniques like Mahalanobis distance, rely heavily on task-specific training and typically require large amounts of in-distribution (ID) data. These methods often define OOD detection in the context of post-processing or retraining models on ID data to differentiate between in- and out-of-distribution samples.

In contrast, LLM-based methods redefine OOD detection by leveraging pre-trained models, allowing them to detect OOD samples without extensive task-specific training. This results in a significant shift in the OOD detection paradigm, moving towards zero-shot and few-shot learning, where models can generalize to new tasks with minimal or no additional training. Key differences include:

- *Performance*: LLM-based methods often outperform traditional methods in zero-shot and few-shot scenarios, where limited labeled data is available. Traditional methods struggle without substantial in-distribution data and retraining.

- *Flexibility*: LLM-based approaches are highly adaptable to new tasks and datasets due to their reliance on vast pre-trained knowledge, while traditional methods require significant retraining for new domains or data types.

- *Efficiency*: Traditional OOD methods rely on explicitly training models with in-distribution data, while LLM-based methods redefine the problem by leveraging pre-existing knowledge in zero-shot or few-shot settings, minimizing the need for retraining or task-specific data preparation.

In conclusion, the transition from traditional OOD detection methods to LLM-based approaches represents a shift from task-specific training and rigid models to more flexible, generalizable sys-

| Method | MVTec AD AUC ↑ | VisA AUC ↑ |
|---|---|---|
| *Zero-shot* | | |
| **CoOp** (Zhou et al., 2022)[†] | 88.8 | 62.8 |
| **CLIP-AC** (Radford et al., 2021)[†] | 71.5 | 65.0 |
| **VAND** (Chen et al., 2023a)[†] | 86.1 | 78.0 |
| **AnomalyCLIP** (Zhou et al., 2024)[†] | **91.5** | 82.1 |
| **CLIP-AD** (Liznerski et al., 2022)[†] | 90.9 | 79.2 |
| **FiLo** (Gu et al., 2024a)[†] | 91.2 | **83.9** |
| *One-shot* | | |
| **SPADE** (Park et al., 2019) | 81.0 | 79.5 |
| **PaDiM** (Defard et al., 2021) | 76.6 | 62.8 |
| **PatchCore** (Roth et al., 2022) | 83.4 | 79.9 |
| **WinCLIP** (Jeong et al., 2023)[†] | 93.1 | 83.8 |
| **AnomalyGPT** (Gu et al., 2024b)[†] | 94.1 | 87.4 |
| **AnomalyDINO-S** (Damm et al., 2024)[†] | **96.6** | **87.4** |

Table 2: Anomaly detection results for MVTec AD and VisA (image-level). Bold indicates the best performance. The methods marked with † are using MLLMs as backbones. The results are are sourced from (Zhou et al., 2024; Gu et al., 2024b)

tems that can handle a wider variety of tasks with minimal additional training. As research continues, a hybrid approach combining the efficiency of traditional methods with the flexibility LLM-based models may offer the most robust solution for diverse OOD detection challenges.

## D Quantitative Analysis and Comparison

While our primary goal is to conduct a systematic literature review of existing methods for anomaly and out-of-distribution (OOD) detection tasks, we acknowledge that including quantitative analysis and comparisons is valuable for understanding the practical implications of these methods.

### D.1 Anomaly Detection

Table 2 presents the quantitative results for image-level anomaly detection on two widely used benchmarks, MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022). The results are sourced from the original papers. We ensured that the experiments across different methods used the same dataset and learning settings for a fair comparison.

### D.2 OOD Detection

Table 3 provides an overview of out-of-distribution (OOD) detection results using ImageNet-1K as the in-distribution (ID) dataset. The methods are evaluated across multiple OOD datasets, including Texture, iNaturalist, Places, and SUN (Van Horn et al., 2018; Zhou et al., 2017; Xiao et al., 2010; Cimpoi et al., 2014), with performance measured using

| Method | Texture | | iNaturalist | | Places | | SUN | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC ↑ | FPR95 ↓ | AUC ↑ | FPR95 ↓ | AUC ↑ | FPR95 ↓ | AUC ↑ | FPR95 ↓ | AUC ↑ | FPR95 ↓ |
| *Traditional posthoc methods* | | | | | | | | | | |
| **MSP** (Hendrycks et al., 2019)[†] | 74.84 | 73.66 | 77.74 | 74.57 | 72.18 | 79.12 | 73.97 | 76.95 | 74.98 | 76.22 |
| **MaxLogit** (Hendrycks et al., 2019)[†] | 88.63 | 48.72 | 88.03 | 60.88 | 87.45 | 55.54 | 91.16 | 44.83 | 88.82 | 52.49 |
| **Energy** (Liu et al., 2020)[†] | 88.22 | 50.39 | 87.18 | 64.98 | 87.33 | 57.40 | 91.17 | 46.42 | 88.48 | 54.80 |
| **ReAct** (Sun et al., 2021)[†] | 88.13 | 49.88 | 86.87 | 65.57 | 87.42 | 56.85 | 91.04 | 46.17 | 88.37 | 54.62 |
| **ODIN** (Liang et al., 2018) [†] | 87.85 | 51.67 | 94.65 | 30.22 | 85.54 | 55.06 | 87.17 | 54.04 | 88.80 | 47.75 |
| *Without Tuning methods* | | | | | | | | | | |
| **MCM** (Ming et al., 2022)[†] | 86.11 | 57.77 | 94.61 | 30.91 | 89.77 | 44.69 | 92.57 | 34.59 | 90.76 | 42.74 |
| **NegLabel** (Jiang et al., 2024) | 90.22 | 43.5 | **99.49** | **1.91** | 91.64 | 35.59 | 95.49 | **20.53** | 94.21 | 25.40 |
| **EOE**  (Cao et al., 2024) | 57.53 | 85.64 | 97.52 | 12.29 | 95.73 | 20.40 | 92.95 | 30.16 | 92.96 | 30.09 |
| *With Tuning methods* | | | | | | | | | | |
| **CoOp** (Zhou et al., 2022) | 89.47 | 45.00 | 93.77 | 29.81 | 90.58 | 40.11 | 93.29 | 40.83 | 91.78 | 51.68 |
| **LoCoOp** (Miyai et al., 2024c)[†] | 90.19 | 42.28 | 96.86 | 16.05 | 91.98 | 32.87 | 95.07 | 23.44 | 93.52 | 28.66 |
| **CLIPN** (Wang et al., 2023)[†] | 90.93 | 40.83 | 95.27 | 23.94 | 92.28 | 33.45 | 93.92 | 26.17 | 93.10 | 31.10 |
| **NegPrompt**  (Nie et al., 2024) | **91.60** | **35.21** | 98.73 | 6.32 | **93.34** | **27.60** | **95.55** | 22.89 | **94.81** | **23.01** |

Table 3: Comprehensive OOD detection results for ImageNet-1K as ID dataset. The black bold indicates the best performance. The results marked with † are sourced from (Wang et al., 2023) and (Miyai et al., 2024c). Others are sourced from their original papers (Wang et al., 2023; Jiang et al., 2024; Cao et al., 2024; Nie et al., 2024)

AUC (Area Under the ROC Curve) and FPR95 (False Positive Rate at 95% True Positive Rate).

# E   Evaluation Metrics

The primary goal of both anomaly detection and out-of-distribution (OOD) detection is to differentiate between normal/in-distribution (ID) samples and abnormal/out-of-distribution (OOD) samples, framing the problem as a binary classification task. Several common metrics are used to evaluate the performance of detectors:

**AUROC (Area Under the Receiver Operating Characteristic curve):** This metric evaluates a detector's overall ability to distinguish between ID or normal and OOD or anomalous samples. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), where:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

Here, TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) correspond to the detector's correct and incorrect classifications.

**AUPR (Area Under the Precision-Recall curve):** The AUPR metric is particularly useful for cases where there is class imbalance, as AUROC can be biased in such situations. The Precision-Recall curve plots precision against recall, where:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

**FPR@N (False Positive Rate at TPR = N%):** This metric evaluates the probability of misclassifying an OOD or anomalous sample as ID or normal when the true positive rate (TPR) is set at a specified value, commonly 90% or 95%. This is crucial for real-world deployments where achieving high accuracy on ID samples is important, while also minimizing false positives for OOD or anomaly detection.

**F1 Score:** The F1 score is a harmonic mean of precision and recall, providing a balanced evaluation of a model's performance across both metrics. It is particularly useful in scenarios where there is an imbalance between the positive and negative classes, as it gives a single metric that reflects both false positives and false negatives.

The F1 score is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the model, while recall measures the proportion of actual positives correctly identified by the model. There are two main variations of the F1 score:

- **Macro F1 Score**: This version computes the F1 score independently for each class (ID and OOD, or normal and anomalous) and then takes the average across classes. It treats all classes equally, making it particularly useful when the class distribution is imbalanced.

- **Micro F1 Score**: This version aggregates the contributions of all classes to calculate the F1

score, considering the total number of true positives, false positives, and false negatives across all classes. It is more sensitive to the performance on the larger class.

A high F1 score indicates that the model maintains a good balance between precision (minimizing false positives) and recall (minimizing false negatives), which is critical in practical OOD detection and anomaly detection tasks.

# F   General Guidelines

We provide general guidelines for selecting appropriate approaches for anomaly or out-of-distribution (OOD) detection, considering key factors such as data modality, efficiency, explanation, and optimization.

**Data Modality:**   The choice of approach is strongly influenced by the type of data being analyzed. For textual data, prompting-based methods may not always offer meaningful interpretations of anomalies or OOD detection, particularly when trying to understand patterns in semantic spaces. In these cases, generating embeddings from LLMs and applying specialized post-hoc detection techniques can lead to better results. Fine-tuning LLMs to produce more relevant embeddings may further enhance detection accuracy. In the case of numerical data, such as time series or tabular data, prompting-based methods have been explored, though they often require carefully designed prompts or fine-tuning to capture the underlying structure of the data. For vision data, including images and videos, the development of multimodal LLMs offers greater flexibility. Both prompting-based and contrasting-based methods can be highly effective, as they are capable of handling the diverse characteristics of multimodal data.

**Efficiency:**   Efficiency is a crucial consideration when choosing between prompting-based and contrasting-based methods. Prompting-based methods can be inefficient for high-precision numerical tasks, as the conversion of numerical values into text results in excessively long input sequences. This inefficiency can become a bottleneck for tasks involving long-term predictions or large datasets, where the computational overhead of generating long outputs becomes significant. In contrast, contrasting-based methods are more efficient for detection tasks for image. By utilizing contrastive objectives to distinguish between positive and nega-

tive samples, these methods excel at zero-shot classification and are computationally more efficient, especially for handling multimodal anomaly and OOD detection. Researchers can also explore the use of contrasting-based approaches to numerical data modalities with their visual representations.

**Explanation:**   In fields where transparency and interpretability are critical, explanation plays a pivotal role in model selection. LLMs offer a unique advantage by not only detecting anomalies but also generating human-like explanations. This capability is especially valuable in domains where actionable insights and interpretable results are essential for decision-making. Research should focus on enhancing the ability of LLMs to explain their outputs in a way that aligns with domain-specific requirements for clarity and transparency.

**Optimization:**   The optimization strategy is highly dependent on the specific data modality and the nature of the detection task. For prompting-based methods, parameter-efficient tuning techniques—such as Low-Rank Adaptation are essential for improving model performance without incurring high computational costs. These techniques enable models to adapt to new tasks efficiently while maintaining their generalization capabilities. This allows the broad generalization capabilities of LLMs to be leveraged while optimizing for specific domain-related tasks, leading to improved accuracy and reduced computational overhead.