Foundation Model-energized Anomaly Detection and Outlier Detection: A Survey

Feng Xiao[1[0000−0002−1410−4295]] and Jicong Fan[⋆2[0000−0001−9665−0355]]

[1] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Guangdong, 518172, P.R. China
fengxiao1@link.cuhk.edu.cn
[2] School of Data Science, The Chinese University of Hong Kong (Shenzhen), Guangdong, 518172, P.R. China
fanjicong@cuhk.edu.cn

Abstract. Anomaly detection (AD), the task of identifying samples that significantly deviate from normative data patterns, is a critical challenge in data analysis and decision-making. It has widespread applications in high-stakes fields such as financial fraud detection, medical diagnosis, and cybersecurity intrusion detection. While AD research has long been active in machine learning, the recent rise of foundation models (FMs) has spurred a surge of new, FM-based techniques. This proliferation of methods provides a diverse toolkit but also creates confusion for learners and practitioners. Consequently, a systematic review is urgently needed to synthesize the existing literature, clarify the current research landscape, and identify persistent challenges. This survey addresses three key aspects of FM-energized anomaly detection: (i) the methodological approaches for leveraging FMs in AD; (ii) the comparative advantages and limitations of FM-based methods against traditional techniques; and (iii) the prevailing trends, thematic preferences, and unresolved challenges in the field. Based on this tripartite analysis, we identify promising yet underexplored research pathways. We anticipate that this survey will not only elucidate current developments and trajectories but also catalyze further innovation within the AD community.

Keywords: Anomaly Detection · Outlier Detection · Foundation Models · Large Language Models (LLMs) · Vision Language Models · Survey.

1 Introduction

Foundation Models (FMs) generally refer to the large-scale neural networks pretrained on extensive and diverse data (often text, image, or both), such as BERT [15], GPT-series models [40,41,10], CLIP [39] and DeepSeek-R1 [19], which can be adapted or fine-tuned to a wide range of downstream tasks. Driven by continuous breakthroughs and unprecedented achievements of FMs on natural language processing (NLP) and computer vision (CV), researchers and practitioners pay more attention to FM-energized machine learning tasks. Anomaly

⋆ Corresponding author

detection, a crucial problem in data science, has numerous applications in high-social-impact fields, such as fraud detection in finance, medical diagnosis in healthcare, fault detection in industry, intrusion detection in cybersecurity, and fatigue detection in transportation. Against this backdrop, a flood of studies on FM-energized anomaly detection has developed in recent years. As shown in Fig. 1, the number of related papers has risen sharply.

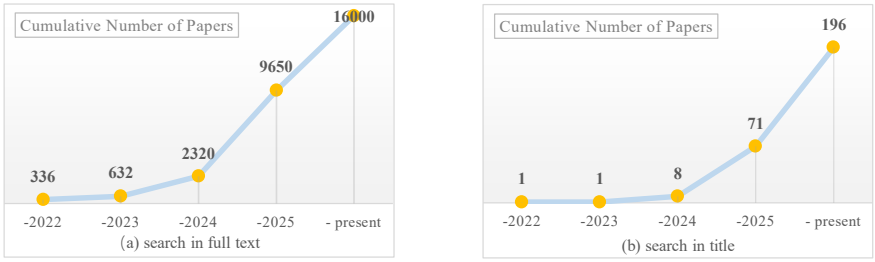


Fig. 1. The cumulative number of papers on anomaly detection associated with foundation models over the years. The statistics are from Google Scholar. The plot (a) shows the search results on full text with keywords '("anomaly detection" OR "outlier detection") AND ("large language models" OR "foundation models")'. The plot (b) shows the search results only on title with keywords 'allintitle:("anomaly detection" OR "outlier detection") (llm OR llms OR "large language models" OR fm OR fms OR "foundation models")'.

The profusion of literature, on one hand, explores the performance of FM-energized techniques on various anomaly detection scenarios and equips us with diverse tools. On the other hand, such abundant methods also cause confusion for learners and practitioners. In response, several recent surveys [47,55,36,42] on FM-based anomaly detection have been introduced to organize and categorize the expanding field. While these reviews provide valuable taxonomies, summaries, and comparisons, they still have respective limitations in addressing such abundant and diverse literature. Specifically, these works, at least, have the following two principal limitations:

- The analysis and comparisons are narrowly focused, often restricted to specific data types or detection scenarios.
- There is a lack of systematic comparison between FM-based and traditional (non-FM-based) anomaly detection methods, leaving their relative advantages and limitations unclear.

This survey aims to address these gaps and provides an effective piece of the puzzle of FM-based anomaly detection, thereby facilitating future research and practical applications of FM-based anomaly detection. The main contributions of this work are summarized as follows.

- This survey introduces a novel taxonomy for the FM-energized anomaly detection methods based on the general anomaly detection process.

- This survey analyzes the advantages and limitations of FM-energized anomaly detection methods compared with traditional (non-FM-based) methods.
- This survey elucidates current prevailing trends, thematic preferences, and unsolved challenges in this field and further identifies promising yet under-explored research pathways for FM-energized anomaly detection.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 provides preliminary knowledge on foundation models and anomaly detection; Section 4 details the novel taxonomy; Section 5 collects useful AD resources including datasets and benchmark literature; Section 6 compares the FM-energized with traditional methods, and analyzes current trends, challenges and possible research problems; and Section 7 concludes this work.

2 Related Surveys

With the rapid development of FM-energized anomaly detection, several considerable surveys have appeared in recent years. Su et al. [47] clearly summarize the prevalent LLM-energized learning paradigms utilized in anomaly detection and prediction tasks and analyze the main problems of the current studies and possible future directions from a high-level perspective. Xu et al. [55] provide a novel taxonomy for anomaly and out-of-distribution (OOD) detection by focusing on how LLMs impact the detection tasks. Following the proposed taxonomy, Xu et al. further discuss the related work under each of the categories and delineate potential challenges and directions for future research. Miyai et al. [36] propose a framework of generalized OOD detection based on Vision Language Models (VLMs). This work provides a comprehensive overview of the field, elucidating key problem settings, established benchmarks, and current methodologies, while also analyzing future research directions for OOD detection in the era of VLMs. Ren et al. [42] review the recent advancements in FM-based anomaly detection and propose a novel taxonomy that classifies these methods into different categories based on their roles in anomaly detection tasks.

Undoubtedly, these works provide in-depth discussion and comparison of current FM-energized or LLM-energized anomaly detection technologies from different facets. All of them analyze the potential challenges and possible directions for future research based on their respective aspects. While considerable contributions, they still have respective limitations in addressing such abundant and diverse literature. In Table 1, we compare the four surveys with ours and categorize them into specific terms to provide a clear preference and concentration of each work. From Table 1, it can be found that the contributions of all the surveys, including ours, are orthogonal and complementary for such a broad and rapidly advancing field. In light of this, in the subsequent sections of this work, we provide clear citations for the issues already thoroughly discussed in the existing reviews, without repeating them.

Table 1. A Comparison among surveys on FM-energized anomaly detection.

Survey	Taxonomy	Models (FMs)	Tasks	Data Types
Su et al. [47]	Utilizing of FMs (e.g. prompt, fine-tuning)	LLMs (e.g. BERT, GPT-2)	Unsupervised AD (clean)	Time series, Log
Xu et al. [55]	Impacting of FMs (detection, generation)	LLMs (e.g. PandaGPT), VLMs (e.g. CLIP)	OOD	N/A
Miyai et al. [36]	fine-tuning of FMs (e.g. zero-shot, few-shot)	VLMs (e.g. CLIP), LVLMs (e.g. GPT-4V, LLaVA)	OOD	Image
Ren et al.[42]	Role of FMs (encoder, detector, interpreter)	LLMs (GPT-2, LLaMa2), VLMs (e.g. CLIP)	N/A	N/A
Ours	Workflow Tree	LLMs, VLMs, LVLMs	Anomaly Detection, Outlier Detection	Log, Time-series, Text, Video, Graph, Image

3 Preliminary Knowledge

3.1 Foundation Models

Foundation Models (FMs) generally refer to the large-scale neural networks pre-trained on extensive and diverse data, which can be adapted or fine-tuned to a wide range of downstream tasks. The term "Foundation Model" was introduced to describe the paradigm shift. Those models, such as BERT [15] and CLIP [39], were no longer just "large language (and/or vision) models" but were becoming the foundation upon which a vast ecosystem of AI applications was being built. This process enables a single model to develop a general-purpose representation of its domain, which can then be specialized for numerous applications. In this work, we regard Foundation Model as a broad category, including early large language (and/or vision) models, like BERT [15], CLIP [39], also including the latest and more powerful large language models (LLMs) and large vision language models (LVLMs), such as DeepSeek-R1 [19] and GPT-4v [37].

3.2 Anomaly Detection

Anomaly detection aims to identify the samples that markedly deviate from the typical patterns. Based on the available training data and different learning paradigms (inductive or transductive), traditional anomaly detection methods [29,43] are categorized into three primary tasks: unsupervised anomaly detection, semi-supervised anomaly detection and outlier detection. We provide formal definitions for each in this section.

Definition 1 (Normal and Abnormal Distribution). *Let $\mathcal{D}_{\mathbf{x}}$ be an unknown bounded normal data distribution on $\mathbb{R}^d$. Define $\mathcal{D}_{\bar{\mathbf{x}}}$ as the complement of $\mathcal{D}_{\mathbf{x}}$ in $\mathbb{R}^d$, namely, $\mathcal{D}_{\mathbf{x}} \oplus \mathcal{D}_{\bar{\mathbf{x}}} = \mathbb{R}^d$. Then $\mathcal{D}_{\bar{\mathbf{x}}}$ is the total abnormal distribution. Let $\mathcal{D}_{\tilde{\mathbf{x}}}$ be an unknown bounded abnormal data distribution on $\mathbb{R}^d$ and $\mathcal{D}_{\tilde{\mathbf{x}}} \subset \mathcal{D}_{\bar{\mathbf{x}}}$.*

Definition 2 (Unsupervised Anomaly Detection (Clean)). *Given Definition 1, let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a set of n samples drawn from distribution $\mathcal{D}_{\mathbf{x}}$ as training data. Then, the goal of the unsupervised AD is to obtain a decision function $f : \mathbb{R}^d \to \{0, 1\}$ by utilizing only $\mathbf{X}$, such that $f(\mathbf{x}_{new}) = 0$ if $\mathbf{x}_{new}$ is drawn from $\mathcal{D}_{\mathbf{x}}$ and $f(\mathbf{x}_{new}) = 1$ if $\mathbf{x}_{new}$ is not drawn from $\mathcal{D}_{\mathbf{x}}$.*

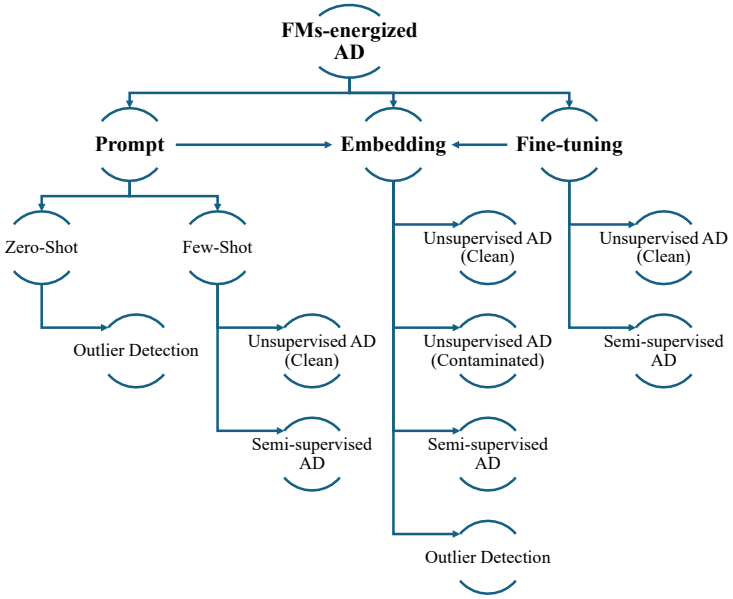


Fig. 2. The corresponding relationship between FM-energized AD methods and the conventional definitions.

Definition 3 (Unsupervised Anomaly Detection (Contaminated)). *Given Definition 1, let* $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$ *be a set of* n *samples drawn from distribution* $\mathcal{D}_{\mathbf{x}}$ *and* $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}\} \subset \mathbb{R}^d$ *be a set of* $\tilde{n}$ *samples drawn from* $\mathcal{D}_{\tilde{\mathbf{x}}}$*, where* $\tilde{n} \ll n$*. Let* $D = \mathbf{X} \cup \tilde{\mathbf{X}}$ *be the training data. Then, the goal of the unsupervised AD is to learn a decision function* $f : \mathbb{R}^d \to \{0, 1\}$ *by utilizing only* D*, such that* $f(\mathbf{x}_{new}) = 0$ *if* $\mathbf{x}_{new}$ *is drawn from* $\mathcal{D}_{\mathbf{x}}$ *and* $f(\mathbf{x}_{new}) = 1$ *if* $\mathbf{x}_{new}$ *is not drawn from* $\mathcal{D}_{\mathbf{x}}$*.*

Definition 4 (Semi-supervised Anomaly Detection). *Given Definition 1, let* $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ *be a set of* n *samples drawn from* $\mathcal{D}_{\mathbf{x}}$ *and* $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}\} \subset \mathbb{R}^d$ *be a set of* $\tilde{n}$ *samples drawn from* $\mathcal{D}_{\tilde{\mathbf{x}}}$*, where* $\tilde{n} \ll n$*. Denote* $y_1 = \cdots = y_n = 0$ *and* $\tilde{y}_1 = \cdots = \tilde{y}_{\tilde{n}} = 1$ *as the labels of normal samples and abnormal samples, respectively. Let* $D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n), (\tilde{\mathbf{x}}_1, \tilde{y}_1), \cdots, (\tilde{\mathbf{x}}_{\tilde{n}}, \tilde{y}_{\tilde{n}})\}$ *be the training data. Semi-supervised AD aims to learn a decision function* $f : \mathbb{R}^d \to \{0, 1\}$ *on* D*, such that* $f(\mathbf{x}_{new}) = 0$ *if* $\mathbf{x}_{new}$ *is drawn from* $\mathcal{D}_{\mathbf{x}}$ *and* $f(\mathbf{x}_{new}) = 1$ *if* $\mathbf{x}_{new}$ *is not drawn from* $\mathcal{D}_{\mathbf{x}}$*.*

Definition 5 (Outlier Detection). *Given Definition 1, let* $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$ *be a set of* n *samples drawn from distribution* $\mathcal{D}_{\mathbf{x}}$ *and* $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{\tilde{n}}\} \subset \mathbb{R}^d$ *be a set of* $\tilde{n}$ *samples drawn from* $\mathcal{D}_{\tilde{\mathbf{x}}}$*. Let* $D = \mathbf{X} \cup \tilde{\mathbf{X}}$ *be the training data. Then, the goal of the outlier detection is to learn a decision function* $f : \mathbb{R}^d \to \{0, 1\}$ *on* D *to identify whether each sample* $\mathbf{x} \in D$ *is from* $\mathcal{D}_{\mathbf{x}}$*.*

When considering FM-energized methods, they often do not have a direct alignment with these conventional definitions. From the perspective of whether

FMs' parameters are updated, the utilization of FMs in anomaly detection can be divided into three categories: embedding-based, prompt-based, and fine-tuning-based, where the first two strategies do not update the parameters of FMs. Fig.2 provides a corresponding relationship between FM-energized AD methods and the conventional definitions.

- The **embedding-based** techniques usually only utilize the FMs to obtain the feature representation of data.
- The **prompt-based** techniques need to restructure the task into a predefined text template that can make full use of the powerful capabilities of language understanding and inference of LLMs.
- The **fine-tuning-based** techniques utilize small-scale and specific datasets to refine the pre-trained FMs to enhance their abilities in a specific domain.

It is worth noting that the aforementioned three categories (embedding-based, prompt-based, and fine-tuning-based) are not an accurate classification for the FM-energized AD literature, since some works would adopt multiple techniques to finish the detection task.

4 FM-energized Anomaly Detection

4.1 Workflow Tree

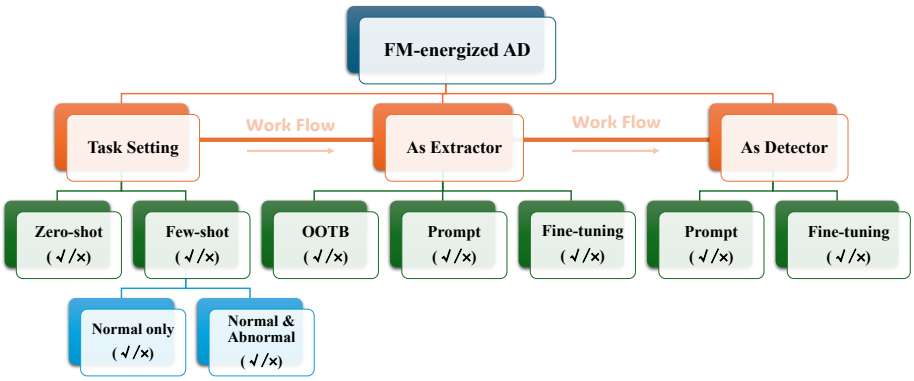


Fig. 3. A comprehensive workflow for each FM-energized AD method.

Through analysis of the aforementioned surveys [47,55,36,42], we found that their taxonomies cannot neatly categorize FM-energized methods into a single category, primarily due to the diversity of anomaly detection tasks and the versatility of foundation models. To better capture these dynamics, we introduce a comprehensive workflow tree (See Fig.3) for FM-energized AD methods by naturally considering the actual detection process, which means that each FM-energized AD method has such a workflow tree to precisely locate within the broader research landscape. In Table 2, we instantiate a collection of workflow trees derived from the FM-energized AD literature, alongside a summary of practical information such as focused data types and the available implementation.

4.2 Task Setting of FM-energized AD

The FM-energized anomaly detection methods do not align well with the conventional definitions (see Section 3.2). Based on few-shot learning, they usually have three settings, including zero-shot, one-shot, and few-shot. In this work, we subsume the one-shot setting under the broader few-shot category.

- In the **zero-shot** setting, FM-energized AD methods are performed on a completely unseen target task. This setting is consistent and comparable with outlier detection (see Definition 5). For example, WinCLIP [23] and AnomalyCLIP [66] followed this setting in their empirical evaluation.
- In the **few-shot** setting, FM-energized AD methods conduct detection guided by a handful of samples from the target task. The utilization of the samples generally includes two ways: (1) fine-tuning the FMs (with parameters updating) such as LLMAD [30]; (2) prompting the FMs (without parameters updating) such as AnomalyGPT [17]. Within the few-shot paradigm, the setting that uses only normal samples corresponds to the unsupervised AD (clean) (See Definition 2) and the setting that uses both normal and abnormal samples aligns with the semi-supervised AD (See Definition 4).

4.3 Using FMs as Extractors

When considering the workflow of anomaly detection within machine learning, a crucial step (often the first step, except for data preprocessing) is feature engineering. The remarkable representational capabilities of FMs establish them as highly effective feature extractors [6]. Based on the literature on FM-energized AD, the utilization of FMs as extractors can be divided into three categories, including OOTB (out of the box), Prompt embedding, and Fine-tuning.

Let $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n\}$ be n original data samples and $\mathrm{FM}(\cdot)$ denotes a feature extractor based a foundation model. The corresponding embedding $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ has

$$\mathbf{x}_i = \begin{cases} \mathrm{FM}(\mathbf{s}_i), & \text{if OOTB;} \\ \mathrm{FM}(\mathrm{Template}(\mathbf{s}_i)), & \text{if Prompt;} \\ \tilde{\mathrm{FM}}(\mathbf{s}_i) \backslash \tilde{\mathrm{FM}}(\mathrm{Template}(\mathbf{s}_i)), & \text{if Fine-tuning.} \end{cases}$$

Where $\mathrm{Template}(\cdot)$ denotes a specific template transformation and the $\tilde{\mathrm{FM}}(\cdot)$ denotes the extractor of the fine-tuned foundation model.

- **OOTB** directly employs FMs as feature extractors without requiring prior fine-tuning and prompt engineering for inputs. Based on this plug-and-play characteristic, the OOTB is particularly prevalent in text and image anomaly detection, such as AD-NLP [7], AD-LLM [57], InCTRL [68], due to the consistent input format for LLMs and VLMs.
- **Prompt embedding** restructures the original samples by a predefined textual template to make full use of the powerful capabilities of language understanding and inference of LLMs. This technique is commonly applied

in industrial anomaly detection (IAD) and non-textual data (e.g. tabular) anomaly detection. For example, AnomalyGPT [17] constructs the templates of image description before feature representation as shown in Fig.4.

– **Fine-tuning** involves updating the parameters of FMs to adapt them to a specific task or domain. When employed as feature extractors for anomaly detection, two fine-tuning strategies can be considered: (1) task fine-tuning that is to refine the FMs by using relevant AD datasets to transform a general model into a specialist; (2) representation fine-tuning that aims to further enhance the representational capabilities of FMs such as Text-ADBench [53] using the fine-tuned LLMs from LLM2Vec [6].

(a) **State-level (normal)**
- c := " [o] "
- c := "flawless [o] "
- c := " perfect [o] "
- c : = "unblemished [o] "
- c : = " [o] without flaw"
- c : = " [o] without defect"
- c : = " [o] without damage"

State-level (anomaly)
- c := " damaged [o] "
- c := " broken [o] "
- c := " [o] with flaw"
- c := " [o] with defect"
- c := " [o] with damage"

(b) **Template-level**
- "a cropped photo of the [c]."
- "a cropped photo of a [c]."
- "a close-up photo of a [c]."
- "a close-up photo of the [c]."
- "a bright photo of a [c]."
- "a bright photo of the [c]."
- "a dark photo of the [c]."
- "a dark photo of a [c]."
- "a jpeg corrupted photo of a [c]."
- "a jpeg corrupted photo of the [c]."
- "a blurry photo of the [c]."
- "a blurry photo of a [c]."
- "a photo of a [c]."
- "a photo of the [c]."
- "a photo of a small [c]."
- "a photo of the small [c]."
- "a photo of a large [c]."
- "a photo of the large [c]."
- "a photo of the [c] for visual inspection."
- "a photo of a [c] for visual inspection."
- "a photo of the [c] for anomaly detection."
- "a photo of a [c] for anomaly detection."

Fig. 4. Image description prompt templates from AnomalyGPT [17]. The complete text can be composed by replacing the token [c] in a template-level text with one of the state-level texts and replacing the token [o] with the object's name.

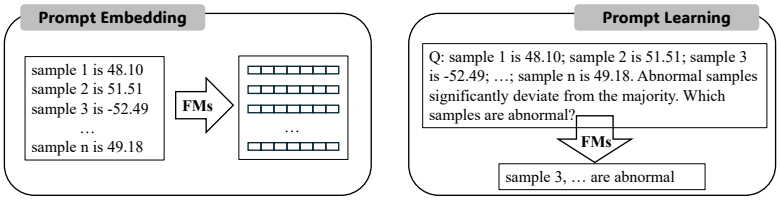
4.4 Using FMs as Detectors

Within the detection stage, two common paradigms for harnessing FMs as detectors are prompt learning (in-context learning) and fine-tuned learning.

– **Prompt learning** does not update the parameters of FMs but only provides task description or a handful of prompted examples to FMs. When only the task description and prompted queries are provided such as TabularAD [26], it corresponds to the setting of zero-shot. When a handful of prompted examples from the target task and queries are provided such as AnomalyGPT [17], it corresponds to the setting of few-shot. It should be noted that the prompt embedding (See Section 4.3) in the stage of feature representation (as an extractor) is different from the prompt learning in the detection stage (as a detector). The former aims to generate embeddings for the prompted samples or auxiliary representations, such as textual descriptions of images. The latter guides the FMs to understand a specific anomaly detection task through textual instructions of the task or labeled examples, leveraging in-context learning to achieve superior detection performance. A clear comparison between the prompt embedding and the prompt learning is illustrated in Fig.5.
– **Fine-tuned learning** involves refining the FMs (updating the parameters) by some small-scale, relevant datasets or the training set of the target task.

Table 2. A summary of the literature on FM-energized Anomaly Detection. The 'MM' and 'F-T' denote 'multimodal' and 'Fine-tuning', respectively. The 'Nor.' and 'Ab.' correspond to 'normal' and 'abnormal', respectively. In the 'Threshold' term, 'Manual' refers to the methods with anomaly scores as the final detection results, and 'Auto' refers to the methods with a binary identification (normal or abnormal).

Methods	Foundation Models							Task Settings				Exp. Settings	
	Model	MM	As Extractor			As Detector		Zero-Shot	Few-Shot		Threshold	Data Type	Code
			OOTB	Prompt	F-T	Prompt	F-T		Nor. only	Nor. & Ab.			
WinCLIP [23] [2023]	CLIP-ViT-B/16+	✓	✓	✓	✗	✗	✗	✓	✓	✗	Manual	Image	link
Xu et al. [56] [2023]	Sentence-BERT	✗	✓	✗	✗	✗	✗	✓	✓	✗	Manual	Text	N/A
Myriad [28] [2023]	EVA-CLIP, Vicuna	✗	✓	✗	✗	✓	✗	✓	✓	✗	Auto	Image	link
Elhafsi et al. [16] [2023]	text-davinci-003	✗	✗	✗	✗	✓	✗	✗	✓	✗	Auto	Text	link
AnomalyGPT [17] [2024]	PandaGPT, Vicuna-7B	✓	✓	✓	✗	✓	✗	✗	✓	✗	Auto	Image	link
AnomalyCLIP [66] [2024]	CLIP-L-14	✓	✓	✓	✗	✗	✗	✓	✗	✗	Manual	Image	link
AnomalyLLM [32] [2024]	Vicuna-7B-v1.5	✗	✗	✓	✗	✗	✗	✗	✗	✓	Manual	Graph	link
TabularAD [26] [2024]	GPT-4	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto	Tabular	N/A
LogFit [2] [2024]	RoBERTa, Longformer	✗	✗	✗	✓	✗	✗	✗	✓	✗	Manual	Log	N/A
ALFA [67] [2024]	CLIP-ViTB/16+, GPT-3.5-turbo-Instruct	✓	✓	✗	✗	✗	✗	✗	✓	✗	Manual	image	N/A
InCTRL [68] [2024]	CLIP-ViT-B/16+	✓	✓	✗	✗	✗	✗	✗	✗	✓	Manual	Image	link
MVFA [21] [2024]	CLIP	✓	✓	✗	✗	✗	✗	✗	✗	✓	Manual	Image	link
AESOP [45] [2024]	many	✓	✓	✗	✗	✗	✗	✗	✓	✗	Manual	Text, Image	link
SIGLLM [3] [2024]	Mistral-7B-Instruct-v0.2, GPT-3.5-turbo-Instruct	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto	Time series	link
SPICED [12] [2024]	GPT-3.5-turbo	✗	✗	✗	✗	✓	✗	✗	✗	✓	Auto	Signal	N/A
Holmes-VAD [60] [2024]	LanguageBind, Vicuna	✓	✓	✗	✗	✓	✓	✗	✗	✓	Auto	Video	link
VAD-LLaMA [34] [2024]	Video-LLaMA	✓	✗	✗	✗	✓	✗	✗	✗	✓	Auto	Video	N/A
LogConfigLocalizer [44] [2024]	GPT-4	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto	Log	N/A
LogiCode [62] [2024]	GPT-4	✗	✗	✗	✗	✓	✗	✗	✗	✓	Auto	Image	link
AnomalyRuler [58] [2024]	CogVLM-17B, GPT-4-1106-Preview, Mistrial-7B-Instruct-v0.2	✗	✗	✗	✗	✓	✗	✗	✓	✗	Auto	Video	link
Audit-LLM [46] [2024]	GPT-3.5-turbo	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto	Log	N/A
NLP-ADBench [27] [2024]	BERT, text-embedding-3-large	✗	✓	✗	✗	✗	✗	✗	✓	✗	Manual	Text	link
AD-LLM [57] [2024]	Llama 3.1 8B Instruct, GPT-4o	✗	✗	✗	✗	✓	✗	✓	✗	✗	Auto	Text	link
Tagfog [13] [2024]	ChatGPT, CLIP	✗	✓	✗	✗	✗	✗	✗	✓	✗	Auto	Image	link
DBAL [18] [2025]	LLaMA 2-chat 13B	✗	✗	✗	✗	✗	✓	✗	✓	✗	Auto	Text	link
LLMAD [30] [2025]	Llama-3-70B, GPT-3.5, GPT-4	✗	✗	✗	✗	✓	✗	✗	✓	✓	Auto	Time-Series	link
TAD-Bench [11] [2025]	many	✗	✓	✗	✗	✗	✗	✗	✓	✗	Manual	Text	N/A
Text-ADBench [53] [2025]	many	✗	✗	✓	✗	✗	✗	✗	✓	✗	Manual	Text	link
AnoLLM [49] [2025]	SmolLM	✗	✗	✗	✗	✗	✓	✗	✓	✗	Manual	Tabular	link
LLM-LADE [63] [2025]	LlaMa3-8B	✗	✗	✗	✗	✗	✓	✗	✗	✓	Auto	Log	link
ICAD-LLM [52] [2025]	Qwen2.5-0.5B	✓	✗	✓	✗	✗	✗	✗	✗	✓	Manual	Time Series, Tabular, Log	link



(a) FMs as extractor (b) FMs as detector

Fig. 5. A naive comparison between prompt embedding and prompt learning when a foundation model is used as an extractor and detector, respectively.

5 Resource Integration

This section provides key resources for anomaly detection research. We catalog the frequently used datasets across different data types and list related benchmark studies in Table 3. These collected resources will support readers in quickly accessing relevant data sources and reviewing the current state of research in specific sub-fields.

5.1 Evaluation Datasets and Benchmark Literature

Table 3 summarizes the frequently used datasets in anomaly detection across seven data types: image, tabular, time-series, text, graph, video, and log.

Table 3. A summary of datasets frequently used in anomaly detection and outlier detection. The 'IAD' refers to 'Industrial Anomaly Detection'.

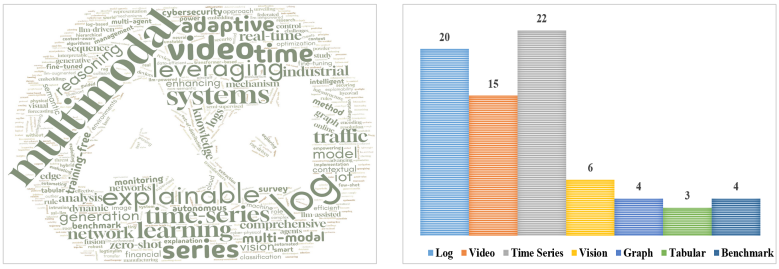
Data Type	Dataset	# Samples	Year	Resource	Benchmark Literature
Image (IAD)	MVTec-AD [9]	5000	2019	link	IM-IAD (2020) [54], MMAD (2025) [25], VIAD (2025) [4]
	BTAD [35]	2830	2021	link	
	MPDD [24]	1346	2021	link	
	MVTec LOCO-AD [8]	3644	2022	link	
	VisA [70]	10821	2022	link	
	AeBAD-S [64]	5570	2023	link	
	GoodsAD [61]	6124	2024	link	
	RIADs [50]	8022	2024	link	
	VAD [5]	5000	2024	link	
Tabular	47 datasets from [20]	N/A	2022	link	ADBench (2022) [20]
Time Series	40 datasets from [31]	N/A	2024	link	TSB-AD (2024) [31], MTSAD (2024) [59], TAB (2025) [38]
Text	8 datasets from [53]	N/A	2025	link	Xu et al (2023). [56], NLP-ADBench (2024) [27], Text-ADBench (2025) [53], AD-LLM (2025) [57], TAD-Bench (2025) [11]
Graph	10 datasets from [48]	N/A	2023	link	GADBench (2023) [48]
Video	Ubnormal [1]	N/A	2022	link	Ubnormal (2022) [1], SmartHome-Bench (2025) [65], RethinkingVAD (2025) [33]
	RethinkingVAD [33]	N/A	2025	link	
	SmartHome-Bench [65]	N/A	2025	link	
Log	19 datasets from [69]	N/A	2023	link	Logeval (2024) [14]

5.2 Evaluation Metrics

During the inference, anomaly detection often can be regarded as a binary classification task. Therefore, the metrics Precision, Recall, F1-score, FPR (False Positive Rate), and FNR (False Negative Rate) can be used to evaluate the performance. All the aforementioned metrics are threshold-dependent and often sensitive to the selection of the threshold. For a holistic measurement of detection performance, the metrics AUROC (Area Under ROC Curve) and AUPRC (Area Under Precision-Recall Curve) are also frequently used to assess the detection performance. There are more detailed descriptions about the evaluation metrics in [47,55].

6 Comparison and Analysis of the Current Research

To provide an overview of current research trends, we visualize the literature titles of FM-energized anomaly detection by word cloud and word frequency in Fig.6 that is based on the search results from Google Scholar with keywords 'allintitle:("anomaly detection" OR "outlier detection") (llm OR llms OR "large language models" OR fm OR fms OR "foundation models")'. In Fig.6, we can easily find that the research on log, video and time-series has received considerable attention. By contrast, the research on graph and tabular data seems to be limited in this setting.



(a) wordcloud of titles (b) partial word frequency of titles

Fig. 6. The wordcloud and word frequency of literature titles of FM-energized AD.

6.1 Advantages of FM-energized Anomaly Detection

Naturally, the key advantages of FM-energized methods include the powerful generalization (fine-tuning learning), the user-friendly design (in-context learning), and the fusion of multimodal data (multimodal learning).

- **Fine-tuning learning** The advent of foundation models has established a new machine learning paradigm: "Pretrained FMs + Task Fine-tuning". In anomaly detection, the learning paradigm requires only one or a few foundation models, adapting them to diverse tasks through fine-tuning or simple, trainable adapters. This paradigm offers a more efficient and sustainable alternative compared to the conventional practice of training a separate model for each dataset.
- **In-context learning** The FMs with chat capabilities (multiple rounds of dialogue), such as GPT-4 [22], DeepSeek-R1 [19], introduce a powerful, training-free mechanism known as in-context learning, which makes the FMs seamless adaptation to new tasks using only natural language instructions (via zero-shot or few-shot prompting). This eliminates the need for extensive task-specific programming and training in real-world applications, enabling a simpler and more efficient strategy. Many works [3,17,26,30] of FM-energized anomaly detection empirically have demonstrated that this approach is feasible and yields good performance in some tasks. Furthermore, the ability to generalize from few-shot or even zero-shot prompts offers an effective solution for data-scarce scenarios, such as rare disease detection.
- **Multimodal learning** Multimodal foundation models offer a transformative advantage for anomaly detection by cross-modal reasoning and contextual understanding, which allows them to identify complex anomalies that are only apparent when synthesizing information from diverse data sources like text, vision, and time-series. Furthermore, a key enhancement of multimodal foundation models is inherent explainability, as these models can generate natural language justifications for their decisions by citing evidence integrated from across different modalities. In high-stakes domains such as financial fraud detection, the interpretability of detection results is paramount.

6.2 Limitations of current FM-energized Anomaly Detection

Although the foregoing analysis paints a promising future for FM-energized anomaly detection, the current research is still at the feasibility testing stage and has, at least, the following limitations.

Table 4. Aggregated numerical results reported in the literature. The FM-energized methods and non-FM-based methods are marked in red and blue, respectively.

Method	Setup	Performance		Method	Setup	Performance	
		AUROC (%)	AUPR (%)			AUROC (%)	AUPR (%)
MVTec-AD (image)				VisA (image)			
WinCLIP [23]	0-shot	91.8	96.5	WinCLIP [23]	0-shot	78.1	81.2
	4-shot	95.2	97.3		4-shot	87.3	88.8
Myriad [28]	0-shot	86.2	-	Myriad [28]	0-shot	78.0	-
	4-shot	96.3	-		4-shot	90.6	-
AnomalyGPT [17]	4-shot	96.3	-	AnomalyGPT [17]	4-shot	90.6	-
AnomalyCLIP [66]	0-shot	91.5	96.2	AnomalyCLIP [66]	0-shot	82.1	85.4
ALFA [67]	0-shot	93.2	97.3	ALFA [67]	0-shot	81.2	84.6
	4-shot	96.5	98.9		4-shot	88.2	89.4
InCTRL [68]	4-shot	94.5	97.2	InCTRL [68]	4-shot	87.7	90.2
	8-shot	95.3	97.7		8-shot	88.7	90.4
PatchCore	full-shot	99.2	99.8	PatchCore	full-shot	95.1	96.2
RD4AD	full-shot	98.6	99.5	RD4AD	full-shot	96.0	96.5
Fraud ecommerce (tabular)				Lymphography (tabular)			
AnoLLM (135M) [49]	full-shot	100.0	99.9	AnoLLM (135M) [49]	full-shot	96.8	85.6
AnoLLM (360M) [49]	full-shot	100.0	97.2	AnoLLM (360M) [49]	full-shot	99.5	93.8
KNN	full-shot	100.0	100.0	KNN	full-shot	86.0	72.0
Iforest	full-shot	50.1	17.2	Iforest	full-shot	67.3	23.2
DeepSVDD	full-shot	100.0	100.0	DeepSVDD	full-shot	89.9	68.0

Method	Setup	Performance		Method	Setup	Performance	
		Best F1 (%)	Delayed F1 (%)			Best F1 (%)	Dealyed F1 (%)
KPI (time series)				Yahoo (time series)			
LLMAD [30]	1-shot (normal)	75.6	30.5	LLMAD [30]	1-shot (normal)	60.2	60.2
	1-shot (abnormal)	76.8	43.1		1-shot (abnormal)	62.2	62.2
	2-shot (1 nor., 1 abnor.)	81.6	64.1		2-shot (1 nor., 1 abnor.)	67.9	65.5
	3-shot (2 nor., 1 abnor.)	84.3	66.7		3-shot (2 nor., 1 abnor.)	72.4	69.5
	5-shot (4 nor., 1 abnor.)	81.2	69.7		5-shot (4 nor., 1 abnor.)	69.1	66.9
Anotransfer	full-shot	68.5	46.1	Anotransfer	full-shot	56.7	49.6
Informer	full-shot	85.9	71.5	Informer	full-shot	70.7	67.1
TFAD	full-shot	75.1	63.1	TFAD	full-shot	77.9	77.5

- **On Detection Performance**
  For anomaly detection, detection performance is the foremost priority. While current FM-energized methods have achieved impressive results with minimal samples (typical fewer than 10), their performance typically remains inferior to state-of-the-art traditional methods trained on a full training set, or related comparison is ignored. We aggregated the related numerical results in Table 4. On image data, the results of FM-energized methods are reported as in their original publications, while those for non-FM-based methods are from a previous benchmark [54]. All tabular data results follow AnoLLM [49], and all time series results follow LLMAD [30]. Although this may seem an unfair comparison for FM-energized methods, it is a pragmatic one for real-world applications. If traditional methods deliver a faster and more accurate inference in spite of training by full-shot, the practical value of FM-energized methods would be limited in many anomaly detection tasks. Notably, the full-shot experiments can be easily conducted in FM-energized AD methods but often are ignored. One possible reason is that increased data volume does not yield significant performance gains for these methods.
- **On Response Speed (Inference Time)**
  The response speed (inference time) is another key metric in anomaly detection, particularly for real-time applications. In this regard, the FM-energized

AD methods currently hold no advantage over many non-FM-energized methods, such as IForest [29], Deep SVDD [43]. Although there are few works [45] that discussed the issue, most current research did not provide related comparison between FM-energized and non-FM-energized AD methods. We argue that such comparison and analysis are essential for quantifying the practical discrepancy between them and will contribute to developing more efficient solutions.

6.3 Possible Research Directions and Problems

Based on the literature review and discussion presented in this work, we have identified several specific directions and problems worthy of further exploration.

- **Insufficient Research on Graph and Tabular data.**
  According to the statistics from Fig.6, the related research on graph [32] and tabular data [26] are relatively insufficient compared with log, video, and time-series. Therefore, the research on graph and tabular data could become more urgent.
- **Unclear Performance Advantage Compared with the non-FM-energized AD methods.**
  This limitation underscores the urgent need for standardized, quantitative benchmarks to assess the true performance difference between current FM-energized and non-FM-energized AD methods. Such benchmarks could accurately answer the following two questions and further reveal more specific research problems.
  1. In which specific AD tasks or domains are FM-energized methods less effective than non-FM-energized methods?
  2. In these specific tasks or domains (based on question 1), how much do these two types of methods differ in terms of detection performance?
- **Slow Response Speed Compared with the non-FM-energized AD methods.**
  This is a general and intractable issue for FM-based methods. In real-time AD applications, response speed is quite crucial to the overall system. Therefore, future research should consider the response speed of FM-energized AD methods and attempt to introduce some efficient inference strategies.
- **Explainability and Trustworthiness.**
  In high-stakes AD scenarios, such as medical diagnosis, the explainability of the identification results and the trustworthiness of the detection process need to be guaranteed to some extent. Leveraging the powerful language understanding, generation, and chain-of-thought (COT) [51] process of LLMs, there are some related studies [60,58] recently. The future research should prioritize generating interpretable justifications, particularly for the identified anomalies.
- **Utilization of Multimodel Data.**
  Multimodal data provide more information and dimensions for the detection process, and several attempts on anomaly detection, such as WinCLIP [23]

AnomalyCLIP [66], have demonstrated their promising potential. This may be a path to outperform the non-FM-energized methods. Therefore, future research should prioritize leveraging multimodal data to address the key challenges, such as scarcity, noise, and class imbalance, thereby enhancing overall detection performance.

7 Conclusion

The advent and advancements of foundation models foster the research on FM-energized anomaly detection. This survey systematically reviews the related literature and aligns it with the conventional definitions of anomaly detection. To synthesize this diverse and rapidly evolving field, we introduce a workflow tree for categorizing methods and integrate key resources, including code, datasets, and benchmark studies. Furthermore, we analyze the advantages of FM-energized anomaly detection from a long run and reveal the limitations of current research. Finally, we outline specific directions and problems to guide future research in FM-energized anomaly detection.

Acknowledgments. The work was partially supported by the General Program of the Natural Science Foundation of Guangdong Province under Grant No.2024A1515011771, the Shenzhen Stability Science Program 2023, and the National Natural Science Foundation of China under Grant No.62376236.

Disclosure of Interests. The authors have no competing interests.

References

1. Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20143–20153 (2022)
2. Almodovar, C., Sabrina, F., Karimi, S., Azad, S.: Logfit: Log anomaly detection using fine-tuned language models. IEEE Transactions on Network and Service Management **21**(2), 1715–1723 (2024)
3. Alnegheimish, S., Nguyen, L., Berti-Equille, L., Veeramachaneni, K.: Large language models can be zero-shot anomaly detectors for time series? arXiv preprint arXiv:2405.14755 (2024)
4. Baitieva, A., Bouaouni, Y., Briot, A., Ameln, D., Khalfaoui, S., Akcay, S.: Beyond academic benchmarks: Critical analysis and best practices for visual industrial anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4015–4025 (2025)
5. Baitieva, A., Hurych, D., Besnier, V., Bernard, O.: Supervised anomaly detection for complex industrial images. In: CVPR (2024)
6. BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., Reddy, S.: Llm2vec: Large language models are secretly powerful text encoders. arXiv preprint arXiv:2404.05961 (2024)

7. Bejan, M., Manolache, A., Popescu, M.: Ad-nlp: A benchmark for anomaly detection in natural language processing. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 10766–10778 (2023)
8. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. International Journal of Computer Vision **130**(4), 947–969 (2022)
9. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
11. Cao, Y., Yang, S., Li, C., Xiang, H., Qi, L., Liu, B., Li, R., Liu, M.: Tad-bench: A comprehensive benchmark for embedding-based text anomaly detection. arXiv preprint arXiv:2501.11960 (2025)
12. Chaudhuri, J., Thapar, D., Chaudhuri, A., Firouzi, F., Chakrabarty, K.: Spiced: Syntactical bug and trojan pattern identification in a/ms circuits using llm-enhanced detection. In: 2024 IEEE Physical Assurance and Inspection of Electronics (PAINE). pp. 1–7. IEEE (2024)
13. Chen, J., Zhang, T., Zheng, W.S., Wang, R.: Tagfog: Textual anchor guidance and fake outlier generation for visual out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1100–1109 (2024)
14. Cui, T., Ma, S., Chen, Z., Xiao, T., Tao, S., Liu, Y., Zhang, S., Lin, D., Liu, C., Cai, Y., et al.: Logeval: A comprehensive benchmark suite for large language models in log analysis. arXiv preprint arXiv:2407.01896 (2024)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
16. Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I.A., Pavone, M.: Semantic anomaly detection with large language models. Autonomous Robots **47**(8), 1035–1055 (2023)
17. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)
18. Guan, W., Cao, J., Gao, J., Zhao, H., Qian, S.: Dabl: Detecting semantic anomalies in business processes using large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11735–11744 (2025)
19. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
20. Han, S., Hu, X., Huang, H., Jiang, M., Zhao, Y.: Adbench: Anomaly detection benchmark. Advances in neural information processing systems **35**, 32142–32159 (2022)
21. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visual-language models for generalizable anomaly detection in medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11375–11385 (2024)

22. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
23. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606–19616 (2023)
24. Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: 2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT). pp. 66–71. IEEE (2021)
25. Jiang, X., Li, J., Deng, H., Liu, Y., Gao, B.B., Zhou, Y., Li, J., Wang, C., Zheng, F.: Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. arXiv preprint arXiv:2410.09453 (2024)
26. Li, A., Zhao, Y., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., Mandt, S.: Anomaly detection of tabular data using llms. arXiv preprint arXiv:2406.16308 (2024)
27. Li, Y., Li, J., Xiao, Z., Yang, T., Nian, Y., Hu, X., Zhao, Y.: Nlp-adbench: Nlp anomaly detection benchmark. arXiv preprint arXiv:2412.04784 (2024)
28. Li, Y., Wang, H., Yuan, S., Liu, M., Zhao, D., Guo, Y., Xu, C., Shi, G., Zuo, W.: Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. arXiv preprint arXiv:2310.19070 (2023)
29. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 eighth ieee international conference on data mining. pp. 413–422. IEEE (2008)
30. Liu, J., Zhang, C., Qian, J., Ma, M., Qin, S., Bansal, C., Lin, Q., Rajmohan, S., Zhang, D.: Large language models can deliver accurate and interpretable time series anomaly detection. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 4623–4634 (2025)
31. Liu, Q., Paparrizos, J.: The elephant in the room: Towards a reliable time-series anomaly detection benchmark. Advances in Neural Information Processing Systems **37**, 108231–108261 (2024)
32. Liu, S., Yao, D., Fang, L., Li, Z., Li, W., Feng, K., Ji, X., Bi, J.: Anomalyllm: Few-shot anomaly edge detection for dynamic graphs using large language models. In: 2024 IEEE International Conference on Data Mining (ICDM). pp. 785–790. IEEE (2024)
33. Liu, Z., Wu, X., Li, W., Yang, L.: Rethinking metrics and benchmarks of video anomaly detection. arXiv preprint arXiv:2505.19022 (2025)
34. Lv, H., Sun, Q.: Video anomaly detection and explanation via large language models. arXiv preprint arXiv:2401.05702 (2024)
35. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: Vt-adl: A vision transformer network for image anomaly detection and localization. In: 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE). pp. 01–06. IEEE (2021)
36. Miyai, A., Yang, J., Zhang, J., Ming, Y., Lin, Y., Yu, Q., Irie, G., Joty, S., Li, Y., Li, H., et al.: Generalized out-of-distribution detection and beyond in vision language model era: A survey. arXiv preprint arXiv:2407.21794 (2024)
37. OpenAI: Gpt-4v(ision) system card (2023)
38. Qiu, X., Li, Z., Qiu, W., Hu, S., Zhou, L., Wu, X., Li, Z., Guo, C., Zhou, A., Sheng, Z., et al.: Tab: Unified benchmarking of time series anomaly detection methods. arXiv preprint arXiv:2506.18046 (2025)

39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
40. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
41. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
42. Ren, J., Tang, T., Jia, H., Xu, Z., Fayek, H., Li, X., Ma, S., Xu, X., Xia, F.: Foundation models for anomaly detection: Vision and challenges. arXiv preprint arXiv:2502.06911 (2025)
43. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. pp. 4393–4402. PMLR (2018)
44. Shan, S., Huo, Y., Su, Y., Li, Y., Li, D., Zheng, Z.: Face it yourselves: An llm-based two-stage strategy to localize configuration errors via logs. In: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. pp. 13–25 (2024)
45. Sinha, R., Elhafsi, A., Agia, C., Foutter, M., Schmerling, E., Pavone, M.: Real-time anomaly detection and reactive planning with large language models. arXiv preprint arXiv:2407.08735 (2024)
46. Song, C., Ma, L., Zheng, J., Liao, J., Kuang, H., Yang, L.: Audit-llm: Multi-agent collaboration for log-based insider threat detection. arXiv preprint arXiv:2408.08902 (2024)
47. Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., Lin, J.: Large language models for forecasting and anomaly detection: A systematic literature review. arXiv preprint arXiv:2402.10350 (2024)
48. Tang, J., Hua, F., Gao, Z., Zhao, P., Li, J.: Gadbench: Revisiting and benchmarking supervised graph anomaly detection. Advances in Neural Information Processing Systems **36**, 29628–29653 (2023)
49. Tsai, C.P., Teng, G., Wallis, P., Ding, W.: Anollm: Large language models for tabular anomaly detection. In: The Thirteenth International Conference on Learning Representations (2025)
50. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22883–22892 (2024)
51. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
52. Wu, Z., Wang, J., Cheng, Z., Zhou, Y., Wang, W., Pu, J., Li, C., Ma, C.: Icad-llm: One-for-all anomaly detection via in-context learning with large language models. arXiv preprint arXiv:2512.01672 (2025)
53. Xiao, F., Fan, J.: Text-adbench: Text anomaly detection benchmark based on llms embedding. arXiv preprint arXiv:2507.12295 (2025)
54. Xie, G., Wang, J., Liu, J., Lyu, J., Liu, Y., Wang, C., Zheng, F., Jin, Y.: Im-iad: Industrial image anomaly detection benchmark in manufacturing. IEEE Transactions on Cybernetics **54**(5), 2720–2733 (2024)
55. Xu, R., Ding, K.: Large language models for anomaly and out-of-distribution detection: A survey. arXiv preprint arXiv:2409.01980 (2024)

56. Xu, Y., Milleret, J., Segond, F.: Comparative analysis of anomaly detection algorithms in text data. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. pp. 1234–1245 (2023)
57. Yang, T., Nian, Y., Li, L., Xu, R., Li, Y., Li, J., Xiao, Z., Hu, X., Rossi, R.A., Ding, K., et al.: Ad-llm: Benchmarking large language models for anomaly detection. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 1524–1547 (2025)
58. Yang, Y., Lee, K., Dariush, B., Cao, Y., Lo, S.Y.: Follow the rules: Reasoning for video anomaly detection with large language models. In: European Conference on Computer Vision. pp. 304–322. Springer (2024)
59. Zhang, C., ZHANG, Y., Wen, Q., Peng, L., Yang, Y., Fan, C., Jiang, M., Fan, L., Sun, L.: Benchmarking multivariate time series anomaly detection with large-scale real-world datasets
60. Zhang, H., Xu, X., Wang, X., Zuo, J., Han, C., Huang, X., Gao, C., Wang, Y., Sang, N.: Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. arXiv preprint arXiv:2406.12235 (2024)
61. Zhang, J., Ding, R., Ban, M., Dai, L.: Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. IEEE Robotics and Automation Letters **9**(3), 2008–2015 (2024)
62. Zhang, Y., Cao, Y., Xu, X., Shen, W.: Logicode: an llm-driven framework for logical anomaly detection. IEEE Transactions on Automation Science and Engineering (2024)
63. Zhang, Z., Li, S., Zhang, L., Ye, J., Hu, C., Yan, L.: Llm-lade: Large language model-based log anomaly detection with explanation. Knowledge-Based Systems **326**, 114064 (2025)
64. Zhang, Z., Zhao, Z., Zhang, X., Sun, C., Chen, X.: Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. Computers in Industry **151**, 103990 (2023)
65. Zhao, X., Zhang, C., Guo, P., Li, W., Chen, L., Zhao, C., Huang, S.: Smarthomebench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3975–3985 (2025)
66. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv preprint arXiv:2310.18961 (2023)
67. Zhu, J., Cai, S., Deng, F., Ooi, B.C., Wu, J.: Do llms understand visual anomalies? uncovering llm's capabilities in zero-shot anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 48–57 (2024)
68. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17826–17836 (2024)
69. Zhu, J., He, S., He, P., Liu, J., Lyu, M.R.: Loghub: A large collection of system log datasets for ai-driven log analytics. In: 2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE). pp. 355–366. IEEE (2023)
70. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European conference on computer vision. pp. 392–408. Springer (2022)