AD-AGENT: A Multi-agent Framework for End-to-end Anomaly Detection

Anonymous ACL submission

Abstract

Anomaly detection (AD) is essential in areas such as fraud detection, network monitoring, and scientific research. However, the diversity of data modalities and the increasing number of specialized AD libraries pose challenges for non-expert users who lack in-depth library-specific knowledge and advanced programming skills. To tackle this, we present AD-AGENT, an LLM-driven multi-agent framework that turns natural-language instructions into fully executable AD pipelines. AD-AGENT coordinates specialized agents for intent parsing, data preparation, library and model selection, documentation mining, and iterative code generation and debugging. Using a shared short-term workspace and a long-term cache, the agents integrate popular AD libraries like PyOD, PyGOD, and TSLib into a unified workflow. Experiments demonstrate that AD-AGENT produces reliable scripts and recommends competitive models across libraries. The system is open-sourced to support further research and practical applications in AD.

1 Introduction and Related Work

Anomaly detection (AD) plays a crucial role in a wide range of applications, including fraud detection (Abdallah et al., 2016), network monitoring (Sun et al., 2023), action recognition (Li et al., 2024b), and medical analysis (Fernando et al., 2021). To handle these diverse data types, the community has released modality-specific open-source libraries that package state-of-the-art models and utilities. Although these libraries accelerate experimentation, each introduces its own data formats and APIs, so users must "juggle" incompatible workflows before they can run even baseline methods. This learning overhead discourages adoption, especially among domain specialists who are not software/data engineers. The stakes are high: Knight Capital lost USD 440 million in 45 minutes when an unchecked trading anomaly cascaded through its systems (Heusser, 2012), and Target's 2013 breach has cost more than 200 million (U.S. Senate Committee on Commerce, Science, and Transportation, 2014). These incidents show that small gaps in an AD pipeline can cause major financial or security failures, showing the need for tooling that is both reliable and easy to integrate.

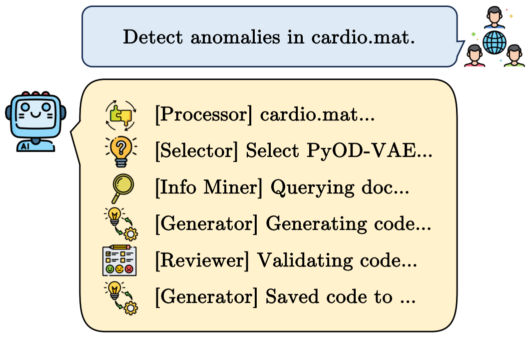

Figure 1: Illustration of AD-AGENT: given a user request, the multi-agent system coordinates each stage to generate a runnable pipeline.

Meanwhile, large language models (LLMs) have demonstrated strong capabilities in reasoning (Guo et al., 2025), code generation (Liu et al., 2023), and tool use (Schick et al., 2023). Recent advances in agent-based systems have further enhanced the potential of LLMs to automate complex, multi-stage tasks that previously required substantial manual effort (Guan et al., 2023) (see extended related work in Appx. A). This presents a compelling opportunity: *Can we develop a general-purpose AD platform that leverages LLMs and existing libraries to build complete detection pipelines from the natural language intents of non-expert users?*

To address this, we introduce AD-AGENT– a multigent framework powered by LLMs that automates the construction of AD pipelines from plain language instructions. It decomposes the AD workflow into specialized agents responsible for user intent interpretation, data processing, library and model selection, knowledge retrieval, code generation and verification, and optional evaluation and tuning. For the memory mechanism, which is the

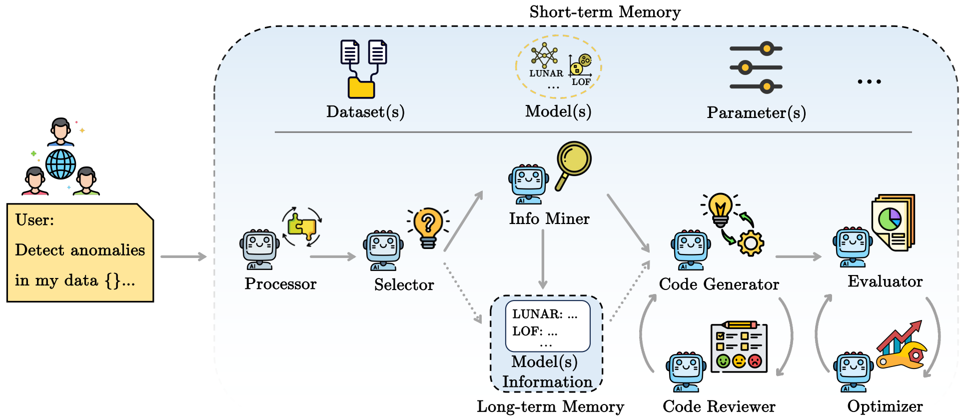

Figure 2: Flowchart of AD-AGENT. Users input natural language instructions and data from various modalities. AD-AGENT coordinates multiple LLM-powered agents via short-term and long-term memory to construct anomaly detection pipelines. Solid arrows represent the default workflow; dashed arrows indicate an optional path that bypasses web searches when algorithm information is stored in long-term memory.

key component to support agent-environment interactions (Zhang et al., 2024), we propose two memories. The short-term shared memory maintains the context of the current session, enabling coordination among agents, while the long-term memory serves as a cache to reduce costly queries across repeated sessions. By combining specialized agents with structured memory, AD-AGENT allows non-expert users to build comprehensive AD pipelines across multiple libraries and modalities using only natural language, relieving the need for library-specific expertise or manual programming. Figure 1 provides an illustration of AD-AGENT.

A survey of prior related LLM-agent work and modality-specific AD libraries is provided in Appendix A. **Our contributions are as follows**:

- **Unified multi-modal-library automation**. We propose the first multi-agent framework that integrates multiple domain-specific AD libraries, enabling end-to-end, cross-modality pipeline construction from natural language.
- **Modular, extensible, and long-lifecycle design**. Loosely coupled agents for reasoning, retrieval, and generation enable AD-AGENT to easily incorporate new libraries and tasks with minimal changes, supporting a long-lasting ecosystem.
- **Accessible to non-experts**. AD-AGENT converts natural language instructions into executable scripts and supports diverse data types, enabling non-expert users without programming skills or specialized knowledge to start easily.
- **Open-source release**. We release AD-AGENT at `https://anonymous.4open.science/r/AD-AGENT-7D26` to provide the community with a practical, extensible platform for LLM-driven AD research and real-world applications.

2 Methodology

We present AD-AGENT, a multi-agent framework that automates AD across diverse modalities and use cases. By integrating established AD libraries – PyOD for multivariate data, PyGOD for graph data, and TSLib for time series – AD-AGENT supports a broad range of models and enables end-to-end automation from user instruction to script.

2.1 Agents

We decompose the detection workflow into multiple subtasks, with each stage handled by a specialized LLM-powered agent, as illustrated in Fig. 2.

Processor. Datasets in practice come in diverse formats (e.g., .csv, .mat, or even natural language), and detection tasks may vary from supervised setups to zero-shot scenarios. The Processor agent serves as the entry point of the system, using LLMs to interpret inputs, infer key attributes (e.g., modality, supervision type), and extract user-specified constraints. It organizes this information into a structured format that guides downstream agents.

Selector. Building on the Processor's output, the Selector agent determines which AD library best aligns with the inferred data modality and task requirements. If the user does not specify a model, the Selector recommends one from the chosen library. Inspired by recent advances in LLM-based model selection (Qin et al., 2025; Chen et al., 2024; Yang et al., 2024), it leverages the LLM's knowledge of models to provide context-aware suggestions tailored to the dataset and task.

Info Miner. Understanding how to apply a model often requires consulting multiple documentation sources, which can be time-consuming and challenging, especially for non-experts. The Info Miner agent performs this background research autonomously. It integrates "Web Search" function from OpenAI (OpenAI, 2025b) to learn from and summarize relevant documents, code examples, and online tutorials. The output includes model descriptions, instructions, and parameter definitions for later code generation.

Code Generator & Reviewer. These two agents collaborate to produce reliable detection scripts. The Generator composes code based on user instructions and knowledge from the Info Miner. To ensure correctness, the Reviewer validates the code through a dry run using LLM-generated synthetic samples, aiming to quickly catch any execution errors. If issues are detected, the two agents enter a feedback loop, iteratively refining the code until a valid and executable pipeline is achieved.

Evaluator & Optimizer. These two agents provide optional extensions for performance evaluation and hyperparameter tuning. The Evaluator runs the pipeline and summarizes detection results when ground truth labels are available for the target dataset. The Optimizer, inspired by Liu et al. (2025), performs LLM-powered hyperparameter tuning based on the provided training dataset. They operate in a feedback loop, iterating between parameter updates and performance assessment.

2.2 Agent Collaboration and Workflow

AD-AGENT facilitates collaboration through two memory structures: a shared **short-term memory** and a persistent **long-term memory**.

The short-term memory serves as the central workspace where agents read and write task-related content. It stores the user input, the processed dataset, selected models, and parameter configurations. This enables agents to operate independently while remaining context-aware.

The long-term memory caches model information retrieved by Info Miner. Since mining from web sources is often time-consuming and resource-intensive, the system first checks this cache for recent summaries before initiating a new web search. It is refreshed periodically (e.g., weekly), allowing the system to benefit from up-to-date resources while avoiding redundant queries.

As shown in Fig. 2, the system begins with the Processor, which interprets the user's input and prepares the data. Based on this context, the Selector determines the appropriate library and, if unspecified by the user, recommends a suitable model. The Info Miner then gathers relevant model details, consulting either the long-term memory or the web.

Table 1: Pipeline generation performance by library, showing success rate (code runs without error), average latency, LLM token usage (input/output), and per-pipeline billing cost in US dollars. The time spent in Reviewer is related to the complexity of models, which explains the increase in TSLib.

Libraries	Success Rate (%)	Time (s)	In/Out Tokens	Cost (US $)
PyOD	100.0	24.0	3,272/667	0.015
PyGOD	91.1	19.6	3,143/673	0.015
TSLib	90.0	125.2	2,680/561	0.012

With this knowledge in place, the Code Generator and Reviewer collaboratively assemble and verify the detection pipeline through an iterative feedback loop until the code is valid and executable. Users may then choose to enable the Evaluator and Optimizer for optional performance assessment and hyperparameter tuning.

This collaborative agent framework allows AD-AGENT to flexibly support multiple data types, including new libraries, adapt to varying input formats, and deliver usable outputs with minimal user effort. Each agent contributes a specialized capability, with LLMs enabling reasoning, adaptation, and coordination across the workflow.

3 Experiments

We evaluate AD-AGENT on reliability and efficiency of the system in constructing executable anomaly detection pipelines from natural language instructions, the quality of model selection, and the effectiveness of long-term memory. For discussion of use cases, see Appx. B.2. Improvements by Optimizer can be found in Appx. B.3.

Datasets and Models. We select datasets and models for each library from their corresponding benchmarks: Chen et al. (2024) for PyOD, Liu et al. (2022) for PyGOD, and Wu et al. (2023) for TSLib. See details in Appx. B.1.

3.1 Pipeline Generation

We first assess whether AD-AGENT can successfully generate runnable pipelines across datasets and models in each supported library. We use *GPT-4o* (OpenAI, 2024) to build all agents in our study. Table 1 presents the success rate, indicating whether the generated code runs without errors, the average generation time, and the average LLM token usage across different dataset–model pairs.

AD-AGENT demonstrates high reliability in producing valid pipelines across modalities, with low latency and manageable cost. We provide a complete example run in Appx. C for reference.

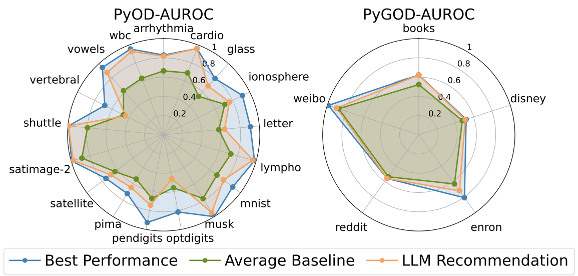

Figure 3: Model selection results for PyOD and PyGOD. We display the average AUROC of models recommended by querying the reasoning LLM three times (duplicates allowed). "Best Performance" marks the highest performance achieved by any available model for each dataset, while "Average Baseline" denotes the mean performance across all available models.

Correction Discussion. The feedback loop between the Code Generator and Reviewer often automatically corrects errors that occur during the initial code generation process. The most frequently fixed issues include missing or incorrectly assigned parameters and incorrect model import names. For example, when the Generator omits a required argument such as `n_features` for DeepSVDD, the Reviewer detects the resulting `TypeError`, references the correct constructor signature via the Info Miner, and amends the script accordingly. These correction cases demonstrate the practical benefit of the collaborative agent loop, allowing AD-AGENT to recover from common errors and increasing the pipeline success rate without user intervention.

Failure Discussion. While AD-AGENT demonstrates high overall reliability, a few recurring failure modes remain. Some failures arise from unaddressed internal data constraints. For instance, GAAN in PyGOD expects binary targets for its loss function, but the pipeline sometimes provides values outside the valid range. This highlights the need for improved data validation and type checking within both the Processor and Generator.

Additionally, some errors stem from library inconsistencies or incorrect functions, such as failed imports of DOMINAT in PyGOD, which is therefore excluded from the experiments, or input-size mismatches for Pyraformer in TSLib with certain datasets. While these are external, they underscore the need for AD-AGENT to integrate version checking and more robust fallback mechanisms.

3.2 Model Selection

We employ *o4-mini* (OpenAI, 2025a) to recommend AD models when the user leaves it unspecified. For each dataset, we query the LLM three times and compute the mean AUROC of selected models. Figure 3 compares the results in PyOD and PyGOD against two baselines: (*i*) the **best** result from any available model, indicating the upper performance limit; and (*ii*) the **average** performance of all available models, representing random selection. See more details and results in Appx. B.4.

Table 2: Average Web Search latency. Long-term memory lookups complete instantly and are omitted.

Libraries	PyOD	PyGOD	TSLib
Time (s)	10.6	12.0	10.8

The LLM's recommendations substantially exceed the average baseline and closely track the best performance in most datasets. This demonstrates that the Selector agent can harness LLM reasoning to choose proper models, simplifying model selection for non-expert users.

3.3 Long-term Memory Efficiency

To quantify the benefit of long-term memory, we compare the Info Miner's lookup latency and cost when using Web Search versus cached summaries. A typical Web Search takes about 10 seconds, as shown in Table 2, and costs 0.035 (US $) per call. In contrast, retrieving the same information from long-term memory is almost instantaneous and incurs no additional cost. This highlights the efficiency of long-term memory.

4 Conclusion

In this work, we introduced AD-AGENT, an LLM-powered multi-agent framework that automates end-to-end AD across multivariate, graph, and time-series data. By decomposing the workflow into specialized agents and coordinating them through short-term and long-term memory, AD-AGENT turns natural language instructions into runnable detection pipelines. Our experiments demonstrate high success rates of the system, accurate model recommendations, and substantial reductions in lookup latency and cost via long-term caching. The system is released for further research.

Future Directions. We plan to: *(i)* broaden AD-AGENT by continually adding new libraries and adapting other data modalities; *(ii)* support conversational interactions so users can iteratively refine pipelines; *(iii)* provide a secure, cloud-based workspace with pre-configured environments to simplify setup; *(iv)* introduce cost-aware planning that balances performance and LLM API budgets; and *(v)* envision a global, community-driven ecosystem where stakeholders collaborate on open-source tools for AD.

Limitations

Despite its flexibility and automation, AD-AGENT has several limitations. The system depends on the accuracy and currency of both the underlying LLMs and external libraries; breaking changes or undocumented features may lead to pipeline failures. Also, not all model or data-specific constraints can be automatically detected, which may result in occasional misconfigurations or runtime errors. Furthermore, AD-AGENT has been validated primarily on standard benchmarks, and its effectiveness and robustness for specialized or proprietary datasets need further systematic investigation.

Ethics Statement

This work adheres to established ethical standards in both research and software development. All experiments are conducted on public datasets, with no personally identifiable or sensitive information processed or disclosed. AD-AGENT is under the BSD 2-clause License, ensuring transparency and reproducibility. The system is designed to assist non-expert users in building AD pipelines. Additionally, ChatGPT was used exclusively to make minor grammatical improvements to the manuscript.

References

Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. 2016. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113.

Roel Bouman, Zaharah Bukhsh, and Tom Heskes. 2024. Unsupervised anomaly detection algorithms on real-world data: how many do we need? *Journal of Machine Learning Research*, 25(105):1–34.

Sihan Chen, Zhuangzhuang Qian, Wingchun Siu, Xingcan Hu, Jiaqi Li, Shawn Li, Yuehan Qin, Tiankai Yang, Zhuo Xiao, Wanghao Ye, and 1 others. 2024. Pyod 2: A python library for outlier detection with llm-powered model selection. *arXiv preprint arXiv:2412.12154*.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2021. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37.

Yile Gu, Yifan Xiong, Jonathan Mace, Yuting Jiang, Yigong Hu, Baris Kasikci, and Peng Cheng. 2025. Argos: Agentic time-series anomaly detection with autonomous rule generation via large language models. *arXiv preprint arXiv:2501.14170*.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. 2022. Adbench: Anomaly detection benchmark. *Advances in neural information processing systems*, 35:32142–32159.

Matthew Heusser. 2012. Software testing lessons learned from knight capital fiasco. *CIO Magazine*. Accessed: 2025-05-16.

Mahnaz Koupaee, Jake W Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, and 1 others. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. *arXiv preprint arXiv:2502.08514*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, and 1 others. 2024a. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. 2024b. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. *arXiv preprint arXiv:2411.08227*.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572.

Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yiling Lou. 2024a. Large language model-based agents for software engineering: A survey. *CoRR*.

Kay Liu, Yingtong Dou, Xueying Ding, Xiyang Hu, Ruitong Zhang, Hao Peng, Lichao Sun, and Philip S Yu. 2024b. Pygod: A python library for graph outlier detection. *Journal of Machine Learning Research*, 25(141):1–9.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, and 1 others. 2022. Bond: Benchmarking unsupervised outlier node detection on static attributed graphs. *Advances in Neural Information Processing Systems*, 35:27021–27035.

Siyi Liu, Chen Gao, and Yong Li. 2025. AgentHPO: Large language model agent for hyper-parameter optimization. In *The Second Conference on Parsimony and Learning (Proceedings Track)*.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2024c. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*.

OpenAI. 2024. Gpt-4o system card.

OpenAI. 2025a. Introducing openai o3 and o4-mini.

OpenAI. 2025b. New tools for building agents.

Yuehan Qin, Yichi Zhang, Yi Nian, Xueying Ding, and Yue Zhao. 2025. MetaOOD: Automatic selection of OOD detection models. In *The Thirteenth International Conference on Learning Representations*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. 2024. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *arXiv preprint arXiv:2408.08902*.

Yongqian Sun, Daguo Cheng, Tiankai Yang, Yuhe Ji, Shenglin Zhang, Man Zhu, Xiao Xiong, Qiliang Fan, Minghan Liang, Dan Pei, and 1 others. 2023. Efficient and robust kpi outlier detection for large-scale datacenters. *IEEE Transactions on Computers*, 72(10):2858–2871.

U.S. Senate Committee on Commerce, Science, and Transportation. 2014. A "kill chain" analysis of the 2013 target data breach. Technical report, Majority Staff Report. Accessed: 2025-05-16.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. 2024. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, and 1 others. 2024. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Appendix: AD-Agent: A Multi-agent Framework for End-to-end Anomaly Detection

A Related Works

LLM-based multi-agent systems have emerged as a powerful paradigm for solving complex tasks through role specialization, planning, and tool use (Guo et al., 2024; Li et al., 2023).

These systems have been successfully applied to domains such as software engineering (Liu et al., 2024a), scientific discovery (Liu et al., 2024c), faithfulness evaluation (Koupaee et al., 2025), and social simulations (Li et al., 2024a). In the context of AD, Audit-LLM (Song et al., 2024) targets insider threat detection through multi-agent coordination, and Argos (Gu et al., 2025) uses LLM agents to generate interpretable anomaly rules for time-series monitoring. While effective, these systems are domain-specific and fixed in scope.

In parallel, several open-source libraries have been developed across different data modalities. Popular libraries such as PyOD (Chen et al., 2024), PyGOD (Liu et al., 2024b), and TSLib (Wang et al., 2024) provide strong support for AD on multivariate, graph, and time series data, respectively. While each library is effective within its domain, they differ in requirements and design. These inconsistencies make integration across libraries non-trivial.

AD-Agent unifies multiple AD libraries within an LLM-driven multi-agent framework.

B Experiments Details

B.1 Datasets and Models

As mentioned in § 3, we adopt datasets and models for each library from corresponding benchmarks.

B.1.1 PyOD

Following PyOD 2 (Chen et al., 2024), we evaluated AD-Agent on 17 widely used datasets originally from ADBench (Han et al., 2022), including arrhythmia, cardio, glass, ionosphere, letter, lympho, mnist, musk, optdigits, pendigits, pima, satellite, satimage-2, shuttle, vertebral, vowels, and WBC. For each dataset, we consider 10 models: ALAD, AnoGAN, AE, AE1SVM, DeepSVDD, DevNet, LUNAR, MO-GAAL, SO-GAAL, and VAE. See more details in Chen et al. (2024).

B.1.2 PyGOD

Following PyGOD (Liu et al., 2024b), we evaluated AD-Agent on 5 real datasets originally from BOND (Liu et al., 2022), including books, disney, enron, reddit, weibo. For each dataset, we consider 9 models: AdONE, ANOMALOUS, AnomalyDAE, CONAD, DONE, GAAN, GUIDE, Radar, and SCAN. See more details in Liu et al. (2022).

Table 3: Detection Performance before and after Optimizer. Better results are highlighted in **bold**.

Models	AUROC(before→ after)	AUPRC(before→ after)
AE	0.7875 → **0.8732**	0.4191 → **0.4959**
ALAD	0.5861 → **0.6103**	0.1454 → **0.1624**
AnoGAN	0.8820 → **0.9438**	0.6050 → **0.7034**
AE1SVM	0.9450 → **0.9779**	0.6748 → **0.8388**
DeepSVDD	0.9259 → **0.9757**	0.6370 → **0.8046**
DevNet	0.0323 → 0.0323	0.0585 → 0.0585
LUNAR	0.5254 → **0.7941**	0.1736 → **0.4462**
MO-GAAL	0.5300 → **0.6200**	0.1900 → **0.2000**
SO-GAAL	0.6687 → **0.7724**	0.3512 → **0.4283**
VAE	0.9800 → 0.9800	0.8300 → 0.8300

B.1.3 TSLib

Wu et al. (2023) presents a benchmark study for TSLib (Wang et al., 2024). Following their approach, we evaluated AD-Agent on 5 real-world datasets from Wu et al. (2023), including MSL, PSM, SMAP, SMD, and SWaT. For each dataset, we consider 10 models: Autoformer, DLinear, ETSformer, FEDformer, Informer, LightTS, Pyraformer, Reformer, TimesNet, and Transformer. See more details in Wu et al. (2023).

B.2 Use Cases Discussion

Our framework supports two common use cases frequently encountered in academic research and real-world deployments.

In research or benchmarking settings, users usually have access to a train/test split and ground-truth anomaly labels for the test set. AD-Agent ingests the training data, builds the model, and reports metrics such as AUROC or F1 on the held-out test set if the user enables the Evaluator. Then the Optimizer can further refine hyperparameters by running an inner loop on the training data and passing a possibly better configuration back to the main pipeline before the final evaluation. This mirrors the evaluation protocol adopted by major AD benchmarks such as ADBench (Han et al., 2022).

In many production scenarios, only one raw, unlabeled dataset is available, and the goal is to identify anomalies directly within this set (Bouman et al., 2024). In this case, AD-Agent detects

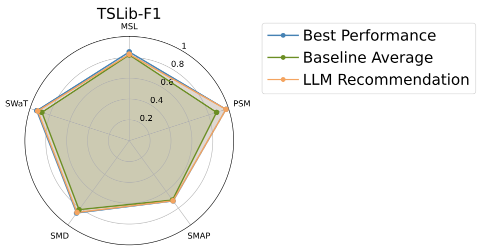

Figure 4: Model selection results for TSLib. We display the average F1-score of models recommended by querying the reasoning LLM three times (duplicates allowed). "Best Performance" marks the highest performance achieved by any available model for each dataset, while "Average Baseline" denotes the mean performance across all available models.

anomalies on the provided data in a single pass; the Evaluator and Optimizer remain inactive unless the user later supplies labels or a separate tuning set.

B.3 Optimizer Improvement

To demonstrate the impact of Optimizer, we evaluated it on the dataset "cardio" within PyOD. As shown in Table 3, Optimizer consistently improved detection quality. These results indicate that the Optimizer agent can automatically refine hyperparameters to produce significantly stronger AD pipelines without human intervention.

B.4 Additional Result of Model Selection

Figure 4 shows the model selection results in TSLib. LLM recommendation outperforms the average baseline in all datasets.

C Example Run

Table 4 presents an actual session of AD-AGENT. In this example, a user requests to run VAE on the "cardio.mat" dataset via a simple natural language command: "Run VAE on cardio.mat." The system interprets the user's intent, processes the data, selects the appropriate library, retrieves model information, and automatically generates a runnable Python script. This example demonstrates the seamless collaboration between agents in AD-AGENT, showing how a single natural language instruction can be transformed into a ready-to-run AD pipeline with minimal user effort.

```
=== [Main] Starting full pipeline ===
=== [Processor] Processing user input ===
Enter command (e.g., 'Run IForest on glass_train.mat and glass_test.mat'):
User: Run VAE on cardio.mat
Experiment Configuration:
   Algorithm: ['VAE']
   Training Dataset: ./data/cardio.mat
   Testing Dataset:
   Parameters:
=== [Processor] User input processing complete
------------------------------------------------------------
=== [Selector] Processing user input ===
=== [Selector] Selecting package & algorithm ===
Package name: pyod
{more content is omitted...}
=== [Selector] Selection complete ===
------------------------------------------------------------
=== [Info Miner] Querying documentation for VAE ===
The Variational Autoencoder (VAE) in PyOD is {more content is omitted...}
**Initialization Function ('__init__'):**
The '__init__' method initializes the VAE model with {more content is omitted...}
**Parameters:**
- 'contamination':
   - **Type**: float in (0., 0.5)
   - **Default**: 0.1
   - **Description**: The proportion of outliers{more content is omitted...}
- 'encoder_neuron_list':
{more content is omitted...}
**Attributes:**
- 'model':
   - **Type**: torch.nn.Module
   - **Description**: The underlying VAE model.
{more content is omitted...}
**Python Dictionary of '__init__' Parameters with Default Values:**
{
   'contamination": 0.1,
   "encoder_neuron_list": [128, 64, 32],
   {more content is omitted...}
}
[Cache Updated] Stored new documentation for VAE
=== [Info Miner] Documentation retrieved for VAE ===
------------------------------------------------------------
=== [Code Generator] Generating code for VAE ===
=== [Code Reviewer] Validating for VAE ===
=== [Code Reviewer] Validation completed for VAE ===
=== [Code Generator] Saved code to ./generated_scripts/VAE_cardio.py ===
{more content is omitted...}
```

Table 4: A real example of AD-AGENT. The user provides a single natural language instruction (highlighted in green), and the system automatically parses the command, retrieves model metadata, and generates an executable Python script. Portions of the printed text are omitted (`{more content is omitted...}`) for brevity.