Novelty Assessment Report

Paper: Scaling Generalist Data-Analytic Agents

PDF URL: https://openreview.net/pdf?id=5PxFqpIYWC

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

Data-analytic agents are emerging as a key catalyst for automated scientific discovery and for the vision of Innovating AI. Current approaches, however, rely heavily on prompt engineering over proprietary models, while open-source models struggle to face diverse-format, large-scale data files and long-horizon, multi-step reasoning that real-world analytics demands. This paper introduces DataMind, a scalable data synthesis and agent training recipe designed to build generalist data-analytic agents. DataMind tackles three key challenges in building open-source data-analytic agents, including insufficient data resources, improper training strategy, and unstable code-based multi-turn rollout. Concretely, DataMind applies 1) a fine-grained task taxonomy and a recursive easy-to-hard task composition mechanism to increase the diversity and difficulty of synthesized queries; 2) a knowledge-augmented trajectory sampling strategy followed by model-based and rule-based filtering; 3) a dynamically adjustable training objective combining both SFT and RL losses; 4) a memory-frugal and stable code-based multi-turn rollout framework. Built on DataMind, we curate DataMind-12K, a high-quality trajectory set spanning diverse domains, task categories, and data file formats for data-analytic tasks. Trained on DataMind-12K, our DataMind-14B achieves state-of-the-art with an average score of 71.16% on multiple data analysis benchmarks, outperforming the strongest proprietary baselines DeepSeek-V3.1 and GPT-5. Our DataMind-7B also performs best among all open-source models with a score of 68.10%. We also incorporate some empirical insights gained from our exploratory trials into the analysis experiments, aiming to provide actionable insights about agentic training for the community. We will release DataMind-12K and DataMind-7B,14B for the community's future research.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Building Generalist Data-Analytic Agents Through Scalable Training**

A total of **28 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Data-Analytic and Query-Driven Agent Systems**
- **Computer-Use and GUI Interaction Agents**
- **Embodied and 3D World Generalist Agents**
- **Game-Based Generalist Agents and World Models**
- **Generalist Agent Frameworks with Minimal Predefinition**
- **Specialized Training Pipelines and Data Synthesis**
- **Domain-Specific Applications and Infrastructure**

Complete Taxonomy Tree

- Building Generalist Data-Analytic Agents Through Scalable Training Survey Taxonomy
- Data-Analytic and Query-Driven Agent Systems
  - Generalist Data-Analytic Agents with Scalable Training ★ (2 papers)
  - [0] Scaling Generalist Data-Analytic Agents (Anon et al., 2026) View paper
  - [9] Agentohana: Design unified data and training pipeline for effective agent learning (Zhang Jian-guo, 2024) View paper
  - Database-Centric Analytics and Query Processing (2 papers)
  - [11] CoddLLM: Empowering Large Language Models for Data Analytics (Zhang Jia-ni, 2025) View paper
  - [16] DaskDB: Scalable Data Science with Unified Data Analytics and In Situ Query Processing (Alex Watson, 2021) View paper
  - Agentic Data Systems for Heterogeneous Data (2 papers)
  - [12] Autonomous data agents: A new opportunity for smart data (Fu, 2025) View paper
  - [24] AgenticData: An Agentic Data Analytics System for Heterogeneous Data (Sun Ji, 2025) View paper
- Computer-Use and GUI Interaction Agents
  - Web-Based GUI Agents (1 papers)
  - [1] GPT-4V(ision) is a Generalist Web Agent, if Grounded (Zheng Boyuan, 2024) View paper
  - General Computer-Use Agents with Compositional Frameworks (2 papers)
  - [3] Agent s2: A compositional generalist-specialist framework for computer use agents (Agashe, 2025) View paper
  - [8] AgentStore: Scalable Integration of Heterogeneous Agents As Specialized Generalist Computer Assistant (Chengyou Jia, 2024) View paper
  - Scalable Task Generation and Trajectory Synthesis for Computer-Use (2 papers)
  - [6] AgentSynth: Scalable Task Generation for Generalist Computer-Use Agents (Xie, 2025) View paper
  - [18] LLMs as Scalable, General-Purpose Simulators For Evolving Digital Agent Training (Wang Yi-ming, 2025) View paper
  - Distributed Data Engines for Computer Agent Training (1 papers)
  - [17] OSGym: Super-Scalable Distributed Data Engine for Generalizable Computer Agents (Zengyi Qin, 2025) View paper

- Embodied and 3D World Generalist Agents
  - 3D World Embodied Agents with Grounding and Reasoning (1 papers)
  - [4] An Embodied Generalist Agent in 3D World (Huang Jiang-yong, 2023) View paper
  - Multi-Embodiment Generalist Policies (2 papers)
  - [2] A Generalist Agent (Reed Scott, 2022) View paper
  - [25] From Multimodal LLMs to Generalist Embodied Agents: Methods and Lessons (Andrew Szot, 2024) View paper
- Game-Based Generalist Agents and World Models
  - World Model-Based Reinforcement Learning Agents (1 papers)
  - [7] Training Agents Inside of Scalable World Models (Hafner, 2025) View paper
  - Foundation Models for Multimodal Game Agents (2 papers)
  - [14] Optimus-3: Towards Generalist Multimodal Minecraft Agents with Scalable Task Experts (Li, 2025) View paper
  - [22] Game-TARS: Pretrained Foundation Models for Scalable Generalist Multimodal Game Agents (Wang Zihao, 2025) View paper
  - Open-Ended Exploration and Skill Learning Benchmarks (2 papers)
  - [19] BuilderBench - A benchmark for generalist agents (Ghugare, 2025) View paper
  - [20] Bootcamp Method for Training General Purpose AI Agents (Vincent Lombardi, 2023) View paper
- Generalist Agent Frameworks with Minimal Predefinition
  - Self-Evolving Agents with Scalable Agentic Reasoning (2 papers)
  - [5] Alita: Generalist Agent Enabling Scalable Agentic Reasoning with Minimal Predefinition and Maximal Self-Evolution (Qiu Jia-Hao, 2025) View paper
  - [15] A Generalist AI Agent SIMA (Mykola Glybovets, 2025) View paper
  - Retrieval-Augmented Generalist Agents for In-Context Adaptation (1 papers)
  - [23] REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments (Sridhar, 2024) View paper
- Specialized Training Pipelines and Data Synthesis
  - Progressive Difficulty Enhancement in Agentic Data Synthesis (1 papers)
  - [28] Synthesizing Agentic Data for Web Agent Training with Progressive Difficulty Enhancement (S Pandit, n.d.) View paper
- Domain-Specific Applications and Infrastructure
  - Retail and Enterprise Agentic AI Platforms (1 papers)
  - [10] Empowering Retail Oss/Bss Platforms With Agentic Ai And Scalable Data Engineering (Motamary, 2025) View paper
  - Multi-Agent Systems for Biological and Clinical Data (1 papers)
  - [21] Multi-Agent AI Systems for Biological and Clinical Data Analysis (J Spieser, 2025) View paper
  - Network Security and Penetration Testing Agents (1 papers)
  - [26] NASimEmu: Network Attack Simulator & Emulator for Training Agents Generalizing to Novel Scenarios (Janisch, 2023) View paper
  - Big Data Infrastructure and Extreme-Scale Computing (2 papers)
  - [13] Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry (M Asch, 2018) View paper
  - [27] Multimodal Approach for Big Data Analytics and Applications (Pal, 2021) View paper

Narrative

Core task: Building generalist data-analytic agents through scalable training. The field of generalist agents has diversified into several distinct branches, each addressing different interaction modalities and problem settings. Data-Analytic and Query-Driven Agent Systems focus on agents that process structured and unstructured data, often interfacing with databases or analytical workflows. Computer-Use and GUI Interaction Agents tackle desktop and web environments, enabling agents to navigate graphical interfaces much like human users. Embodied and 3D World Generalist Agents extend capabilities into spatial reasoning and physical simulation, while Game-Based Generalist Agents and World Models leverage interactive game environments to train versatile policies. Generalist Agent Frameworks with Minimal Predefinition emphasize open-ended architectures that avoid task-specific engineering, and Specialized Training Pipelines and Data Synthesis explore methods for generating diverse training data at scale. Domain-Specific Applications and Infrastructure round out the taxonomy by addressing deployment challenges in real-world sectors such as retail or biology.

Within this landscape, a particularly active line of work centers on scalable training regimes that combine synthetic data generation with iterative refinement. For instance, AgentSynth[6] and Agentohana[9] illustrate how large-scale data synthesis can bootstrap agent capabilities, while Agent s2[3] and Alita[5] explore different strategies for balancing generalization with task-specific fine-tuning. Scaling Data-Analytic Agents[0] sits squarely in the Data-Analytic and Query-Driven branch, emphasizing scalable training pipelines tailored to analytical workflows. Compared to Agentohana[9], which also targets data-centric tasks, Scaling Data-Analytic Agents[0] places greater emphasis on end-to-end scalability and the integration of diverse data modalities. This positioning highlights an ongoing tension in the field: whether to pursue broad generalist frameworks or to specialize training pipelines for high-stakes analytical domains, a question that remains central to advancing both robustness and practical deployment.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Agentohana: Design unified data and training pipeline for effective agent learning

Authors: Zhang Jian-guo, Jianguo Zhang, Lan Tian, Tian Lan, Murthy, et al. (40 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Autonomous agents powered by large language models (LLMs) have garnered significant research attention. However, fully harnessing the potential of LLMs for agent-based tasks presents inherent challenges due to the heterogeneous nature of diverse data sources featuring multi-turn trajectories. In this paper, we introduce \textbf{AgentOhana} as a comprehensive solution to address these challenges. \textit{AgentOhana} aggregates agent trajectories from distinct environments, spanning a wide array o...

Relationship Analysis

Both papers belong to the same taxonomy category of generalist data-analytic agents trained through scalable methods, sharing a focus on building open-source agents via trajectory-based training and data synthesis. They overlap in addressing the challenges of insufficient training data, diverse task coverage, and multi-turn agent interactions for data-analytic tasks. However, DATAMIND focuses specifically on data analysis with a fine-grained task taxonomy, recursive query composition, and code-based execution for analytical tasks, while AgentOhana emphasizes unified data pipeline design across heterogeneous agent environments (web navigation, tool usage, APIs) with a standardized multi-turn trajectory format and AgentRater filtering mechanism.

Contributions Analysis

Overall novelty summary. DataMind proposes a scalable data synthesis and agent training recipe for building generalist data-analytic agents, addressing challenges in open-source model development for diverse-format data and long-horizon reasoning. The paper resides in the 'Generalist Data-Analytic Agents with Scalable Training' leaf, which contains only two papers total. This sparse population suggests the specific combination of scalable training pipelines and generalist data analytics remains relatively underexplored, positioning the work in an emerging rather than saturated research direction within the broader taxonomy of agent systems.

The taxonomy reveals neighboring research directions that contextualize DataMind's positioning. Database-Centric Analytics focuses on SQL and structured query processing, while Agentic Data Systems emphasizes autonomous analysis of heterogeneous sources. Computer-Use Agents tackle GUI interaction rather than data analytics, and Specialized Training Pipelines explore progressive difficulty enhancement for web or research agents. DataMind bridges scalable training methodology with data-analytic capabilities, diverging from database-specific systems by targeting diverse data formats and from computer-use agents by emphasizing analytical reasoning over interface manipulation.

Among thirty candidates examined, none clearly refuted any of DataMind's three core contributions: the scalable training recipe, the DataMind-12K trajectory dataset, or the resulting 7B/14B models. Each contribution was assessed against ten candidates with zero refutable overlaps identified. This suggests that within the limited search scope, the specific combination of fine-grained task taxonomy, knowledge-augmented trajectory sampling, hybrid SFT-RL training, and memory-frugal rollout appears relatively novel. However, the analysis explicitly covers top-K semantic matches rather than exhaustive prior work, leaving open the possibility of relevant work outside this candidate pool.

Based on the limited literature search of thirty candidates, DataMind appears to occupy a sparsely populated research direction with no clear prior work overlap detected. The taxonomy structure confirms that scalable training for generalist data-analytic agents remains less crowded than adjacent areas like computer-use or game-based agents. These signals suggest meaningful novelty within the examined scope, though the analysis does not claim exhaustive coverage of all potentially relevant prior work in data analytics or agent training.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: DATAMIND scalable data synthesis and agent training recipe

Description: The authors propose DATAMIND, a comprehensive pipeline that addresses key challenges in building open-source data-analytic agents through fine-grained task taxonomy, recursive task composition, knowledge-augmented trajectory sampling, dynamically adjustable training objectives combining SFT and RL losses, and a memory-frugal code-based multi-turn rollout framework.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. On the diversity of synthetic data and its impact on training large language models

URL: View paper

Brief Assessment

Synthetic Data Diversity[32] focuses on measuring diversity in synthetic pre-training data for general LLMs using clustering metrics, not on building data-analytic agents with fine-grained task taxonomies, trajectory sampling, or code-based multi-turn rollout frameworks as in DATAMIND.

2. Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continue Pre-training of Large Language Models

URL: View paper

Brief Assessment

Synthesize-on-Graph[34] focuses on synthetic data generation for continued pre-training of LLMs using cross-document knowledge graphs, not on building data-analytic agents with multi-turn code execution and RL-based training pipelines.

3. TARGA: Targeted Synthetic Data Generation for Practical Reasoning over Structured Data

URL: View paper

Brief Assessment

TARGA[30] focuses on semantic parsing for KBQA/text2sql tasks through targeted synthetic query generation, not on building generalist data-analytic agents with multi-turn code execution and RL training as in the original paper.

4. CoddLLM: Empowering Large Language Models for Data Analytics

URL: View paper

Brief Assessment

CoddLLM[11] focuses on a different domain (general data analytics with SQL and table selection) rather than the specific data-analytic agent training recipe proposed in the original paper. The candidate addresses data analytics tasks but does not present a comparable pipeline for training generalist data-analytic agents with the specific components described in the original contribution.

5. Mag-v: A multi-agent framework for synthetic data generation and verification

URL: View paper

Brief Assessment

Mag-v[31] focuses on synthetic question generation for testing agents and trajectory verification using multi-agent systems, not on comprehensive data-analytic agent training recipes combining fine-grained task taxonomy, recursive composition, knowledge-augmented sampling, dynamic SFT+RL objectives, and memory-frugal rollout frameworks as proposed in the original paper.

6. Repurposing synthetic data for fine-grained search agent supervision

URL: View paper

Brief Assessment

Repurposing Synthetic Data[33] focuses on repurposing entity information from synthetic data for fine-grained reward signals in search agents, not on comprehensive data synthesis pipelines for data-analytic agents. The candidate addresses reward formulation in RL for web search tasks, while the original contribution encompasses task taxonomy, recursive composition, knowledge-augmented sampling, and code-based rollout frameworks for data analysis.

7. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning

URL: View paper

Brief Assessment

Ui-tars-2[36] focuses on GUI agent training with multi-turn RL in interactive environments (browsers, VMs, games), while DATAMIND targets data-analytic agents that process tabular/database files via code generation. The domains, task types, and training objectives differ fundamentally.

8. AgentSynth: Scalable Task Generation for Generalist Computer-Use Agents

URL: View paper

Brief Assessment

AgentSynth[6] focuses on synthesizing computer-use tasks for GUI agents in desktop environments, while the original paper addresses data-analytic agents that process data files through code generation. These are distinct agent domains with different challenges and methodologies.

9. Robocasa: Large-scale simulation of everyday tasks for generalist robots

URL: View paper

Brief Assessment

Robocasa[29] focuses on simulation frameworks for robotic manipulation in kitchen environments, not on data-analytic agents or training recipes for data analysis tasks. The domains and technical approaches are fundamentally different.

10. Synthetic Data RL: Task Definition Is All You Need

URL: View paper

Brief Assessment

Synthetic Data RL[35] focuses on general task adaptation using synthetic data for RL training across diverse domains (math, medicine, law, finance), not specifically on building data-analytic agents with code-based multi-turn rollout frameworks and fine-grained task taxonomies for data analysis tasks.

Contribution 2: DATAMIND-12K high-quality trajectory dataset

Description: The authors create DATAMIND-12K, a curated training dataset that covers diverse task categories and data file formats for data analysis tasks, enabling the training of generalist data-analytic agents.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Deep learning for trajectory data management and mining: A survey and beyond

URL: View paper

Brief Assessment

Trajectory Deep Learning[37] focuses on trajectory data management and mining in mobility/transportation contexts, not on data-analytic agent training datasets. The candidate discusses trajectory datasets for spatial-temporal analysis, while the original paper presents a dataset for training generalist data-analytic agents across diverse task categories and file formats.

2. On collaborative multi-UAV trajectory planning for data collection

URL: View paper

Brief Assessment

Multi-UAV Trajectory Planning[41] focuses on UAV trajectory optimization for IoT data collection in physical environments, not on creating trajectory datasets for training data-analytic agents. The domains and applications are fundamentally different.

3. Trajectory generation: a survey on methods and techniques

URL: View paper

Brief Assessment

Trajectory Generation Survey[44] focuses on trajectory generation methods for mobility and spatial data (e.g., next-location prediction, diffusion models for trajectory data), not on creating datasets of reasoning trajectories for data-analytic agent training tasks.

4. OmniWorld: A Multi-Domain and Multi-Modal Dataset for 4D World Modeling

URL: View paper

Brief Assessment

OmniWorld[39] focuses on multi-modal 4D world modeling data (RGB, depth, camera poses, optical flow) for geometric reconstruction and video generation tasks, not data-analytic agent trajectories or code-based reasoning for data analysis.

5. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset

URL: View paper

Brief Assessment

Schema-Guided Dialogue[38] focuses on multi-domain conversational dialogue for virtual assistants, not data-analytic trajectory datasets. The domains and task types are fundamentally different from data analysis tasks.

6. OnSiteVRU: A High-Resolution Trajectory Dataset for High-Density Vulnerable Road Users

URL: View paper

Brief Assessment

OnSiteVRU[42] focuses on physical trajectory data for vulnerable road users (pedestrians, cyclists) in traffic scenarios, not data-analytic task trajectories for training AI agents to perform data analysis.

7. Trajectory design for UAV-based Internet of Things data collection: A deep reinforcement learning approach

URL: View paper

Brief Assessment

UAV Trajectory DRL[45] focuses on UAV flight trajectory optimization for IoT data collection in 3D urban environments, not on creating training datasets for data-analytic agents. The trajectories discussed are physical movement paths for unmanned aerial vehicles, completely different from the reasoning/code execution trajectories in DATAMIND-12K for data analysis tasks.

8. A federated pedestrian trajectory prediction model with data privacy protection

URL: View paper

Brief Assessment

Federated Pedestrian Prediction[40] focuses on pedestrian trajectory prediction in physical spaces with privacy-preserving federated learning, not on creating datasets for data-analytic tasks across diverse domains and file formats.

9. Trajectory Data Collection with Local Differential Privacy

URL: View paper

Brief Assessment

Trajectory Local Privacy[43] focuses on privacy-preserving trajectory data collection using local differential privacy mechanisms, not on creating training datasets for data-analytic agents. The term 'trajectory' refers to location movement paths, not agent solution trajectories for data analysis tasks.

10. MATRIX: multi-agent trajectory generation with diverse contexts

URL: View paper

Brief Assessment

MATRIX[46] focuses on generating multi-agent human navigation trajectories for robotics applications, not data-analytic task trajectories spanning diverse domains and file formats as in DATAMIND-12K.

Contribution 3: DATAMIND-7B and DATAMIND-14B state-of-the-art models

Description: The authors develop DATAMIND-7B and DATAMIND-14B models trained on their curated dataset, achieving state-of-the-art performance on data analysis benchmarks and outperforming both proprietary and open-source baselines.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. ChatGPT vs state-of-the-art models: a benchmarking study in keyphrase generation task

URL: View paper

Brief Assessment

ChatGPT Keyphrase Benchmark[51] focuses on keyphrase generation tasks in natural language processing, not data-analytic agents or reinforcement learning frameworks for data analysis.

2. Predictive performance of presenceâ□□only species distribution models: a benchmark study with reproducible code

URL: View paper

Brief Assessment

Species Distribution Benchmark[56] focuses on species distribution modeling using presence-only ecological data, not data-analytic agents or general data analysis benchmarks. The domains, tasks, and model architectures are fundamentally different.

3. Exploring large language models for qualitative data analysis

URL: View paper

Brief Assessment

LLMs Qualitative Analysis[55] focuses on qualitative data analysis workflows (document classification, information extraction, span classification, text generation) rather than developing state-of-the-art data-analytic agent models. The candidate evaluates existing open-source LLMs (Llama 3.1, Gemma 2, Mistral Nemo) on QDA tasks but does not claim to develop new SOTA models for data analysis benchmarks.

4. Simpo: Simple preference optimization with a reference-free reward

URL: View paper

Brief Assessment

Simpo[49] focuses on preference optimization algorithms for aligning LLMs with human feedback, not on developing state-of-the-art data analysis models or benchmarks. The candidate addresses a fundamentally different problem domain (RLHF/DPO optimization) compared to the original's data-analytic agent training.

5. When do neural nets outperform boosted trees on tabular data?

URL: View paper

Brief Assessment

Neural Nets Tabular[48] focuses on comparing neural networks versus gradient-boosted decision trees on tabular data across 176 datasets, not on developing state-of-the-art data-analytic agent models for multi-step reasoning tasks.

6. Ultrafeedback: Boosting language models with high-quality feedback

URL: View paper

Brief Assessment

Ultrafeedback[52] focuses on constructing preference datasets and training reward/critique models for RLHF, not on developing state-of-the-art data-analytic agents that achieve top performance on data analysis benchmarks.

7. Evaluating vision and pathology foundation models for computational pathology: a comprehensive benchmark study

URL: View paper

Brief Assessment

Vision Pathology Benchmark[50] focuses on benchmarking foundation models for computational pathology across histopathological datasets, not on developing state-of-the-art models for general data analysis tasks.

8. Qlora: Efficient finetuning of quantized llms

URL: View paper

Brief Assessment

Qlora[54] focuses on efficient finetuning methods for quantized LLMs, not on developing state-of-the-art data analysis models or benchmarking on data analysis tasks.

9. The cell tracking challenge: 10 years of objective benchmarking

URL: View paper

Brief Assessment

Cell Tracking Challenge[47] focuses on benchmarking cell segmentation and tracking algorithms in microscopy videos, not on developing general data-analytic agents or achieving state-of-the-art performance on data analysis benchmarks.

10. Evaluating the Performance of Large Language Models on GAOKAO Benchmark

URL: View paper

Brief Assessment

GAOKAO Benchmark[53] evaluates LLMs on Chinese college entrance examination questions across multiple subjects, not on data analysis benchmarks. The candidate focuses on educational assessment tasks rather than data-analytic agent capabilities.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Scaling Generalist Data-Analytic Agents View paper
- [1] GPT-4V(ision) is a Generalist Web Agent, if Grounded View paper
- [2] A Generalist Agent View paper
- [3] Agent s2: A compositional generalist-specialist framework for computer use agents View paper
- [4] An Embodied Generalist Agent in 3D World View paper
- [5] Alita: Generalist Agent Enabling Scalable Agentic Reasoning with Minimal Predefinition and Maximal Self-Evolution View paper
- [6] AgentSynth: Scalable Task Generation for Generalist Computer-Use Agents View paper
- [7] Training Agents Inside of Scalable World Models View paper
- [8] AgentStore: Scalable Integration of Heterogeneous Agents As Specialized Generalist Computer Assistant View paper
- [9] Agentohana: Design unified data and training pipeline for effective agent learning View paper
- [10] Empowering Retail Oss/Bss Platforms With Agentic Ai And Scalable Data Engineering View paper
- [11] CoddLLM: Empowering Large Language Models for Data Analytics View paper
- [12] Autonomous data agents: A new opportunity for smart data View paper
- [13] Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry View paper
- [14] Optimus-3: Towards Generalist Multimodal Minecraft Agents with Scalable Task Experts View paper
- [15] A Generalist AI Agent SIMA View paper
- [16] DaskDB: Scalable Data Science with Unified Data Analytics and In Situ Query Processing View paper
- [17] OSGym: Super-Scalable Distributed Data Engine for Generalizable Computer Agents View paper
- [18] LLMs as Scalable, General-Purpose Simulators For Evolving Digital Agent Training View paper
- [19] BuilderBench - A benchmark for generalist agents View paper
- [20] Bootcamp Method for Training General Purpose AI Agents View paper
- [21] Multi-Agent AI Systems for Biological and Clinical Data Analysis View paper
- [22] Game-TARS: Pretrained Foundation Models for Scalable Generalist Multimodal Game Agents View paper
- [23] REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments View paper
- [24] AgenticData: An Agentic Data Analytics System for Heterogeneous Data View paper
- [25] From Multimodal LLMs to Generalist Embodied Agents: Methods and Lessons View paper
- [26] NASimEmu: Network Attack Simulator & Emulator for Training Agents Generalizing to Novel Scenarios View paper
- [27] Multimodal Approach for Big Data Analytics and Applications View paper
- [28] Synthesizing Agentic Data for Web Agent Training with Progressive Difficulty Enhancement View paper
- [29] Robocasa: Large-scale simulation of everyday tasks for generalist robots View paper
- [30] TARGA: Targeted Synthetic Data Generation for Practical Reasoning over Structured Data View paper
- [31] Mag-v: A multi-agent framework for synthetic data generation and verification View paper
- [32] On the diversity of synthetic data and its impact on training large language models View paper
- [33] Repurposing synthetic data for fine-grained search agent supervision View paper
- [34] Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continue Pre-training of Large Language Models View paper
- [35] Synthetic Data RL: Task Definition Is All You Need View paper
- [36] Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning View paper
- [37] Deep learning for trajectory data management and mining: A survey and beyond View paper
- [38] Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset View paper
- [39] OmniWorld: A Multi-Domain and Multi-Modal Dataset for 4D World Modeling View paper
- [40] A federated pedestrian trajectory prediction model with data privacy protection View paper
- [41] On collaborative multi-UAV trajectory planning for data collection View paper
- [42] OnSiteVRU: A High-Resolution Trajectory Dataset for High-Density Vulnerable Road Users View paper
- [43] Trajectory Data Collection with Local Differential Privacy View paper
- [44] Trajectory generation: a survey on methods and techniques View paper
- [45] Trajectory design for UAV-based Internet of Things data collection: A deep reinforcement learning approach View paper
- [46] MATRIX: multi-agent trajectory generation with diverse contexts View paper
- [47] The cell tracking challenge: 10 years of objective benchmarking View paper
- [48] When do neural nets outperform boosted trees on tabular data? View paper
- [49] Simpo: Simple preference optimization with a reference-free reward View paper

- [50] Evaluating vision and pathology foundation models for computational pathology: a comprehensive benchmark study View paper
- [51] ChatGPT vs state-of-the-art models: a benchmarking study in keyphrase generation task View paper
- [52] Ultrafeedback: Boosting language models with high-quality feedback View paper
- [53] Evaluating the Performance of Large Language Models on GAOKAO Benchmark View paper
- [54] Qlora: Efficient finetuning of quantized llms View paper
- [55] Exploring large language models for qualitative data analysis View paper
- [56] Predictive performance of presenceâ□□only species distribution models: a benchmark study with reproducible code View paper