Preprints.org

Brief Report

Not peer-reviewed version

Trustworthy Multi-Agent Systems: A Human-in-the-Loop Framework for Monitoring, Governance, and Adaptive Correction

Shravani Mahale *

Posted Date: 2 November 2026

Keywords: multi-agent systems; trustworthy AI; human-in-the-loop; autonomous agents; AI governance; explainable AI; adaptive correction; auditability; safe mode; trust monitoring

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Brief Report

Trustworthy Multi-Agent Systems: A Human-in-the-Loop Framework for Monitoring, Governance, and Adaptive Correction

Shravani Mahale

Department of AIDS, Sanjivani University, Kopargaon, Maharashtra, India; shravanimahale@gmail.com

Abstract

Autonomous multi-agent systems are increasingly being designed to perform complex tasks with limited human intervention. However, increasing autonomy introduces challenges related to trust, transparency, accountability, controllability, and safe decision-making. This paper presents a web-based control-room prototype for a Trustworthy Multi-Agent System designed to monitor and govern autonomous agents while maintaining human oversight. The proposed framework integrates trust monitoring, evidence-based decision-making, adaptive self-correction, auditability, explainability, runtime monitoring, and Safe Mode controls. The system provides a Trust Posture dashboard for monitoring overall system trust, agent activity, self-corrections, human interventions, and trust trajectories. An Agent Roster provides information about individual agents, including their roles, trust scores, operational status, run counts, and available operator controls. Recent audit events maintain a human-readable record of approvals, reviews, corrections, and evidence updates. A Safe Mode mechanism provides controlled autonomy by requiring human confirmation before new autonomous actions can proceed. The evaluation framework considers Trust Score, Confidence Level, Agent Activity, Self-Correction Count, Human Intervention Count, Agent Status, Audit Events, and Decision/Action Status as measurable variables. Purposive sampling is used to identify meaningful system activities for evaluation. The proto-type demonstrates how monitoring, human oversight, evidence-based decisions, and adaptive correction can be integrated into a unified governance framework for trustworthy autonomous multi-agent systems.

Keywords: multi-agent systems; trustworthy AI; human-in-the-loop; autonomous agents; AI governance; explainable AI; adaptive correction; auditability; safe mode; trust monitoring

I. Introduction

The rapid development of autonomous artificial intelligence has resulted in systems capable of performing tasks with increasing levels of independence. Multi-agent systems extend this capability by allowing multiple autonomous agents to perform specialized tasks and interact within a common environment. Although such systems can improve automation and efficiency, increased autonomy creates challenges concerning trust, accountability, transparency, explainability, and human control.

A trustworthy autonomous system should not only perform tasks successfully but should also provide mechanisms through which its actions can be monitored, evaluated, corrected, and reviewed. The Trustworthy Multi-Agent System presented in this work addresses these requirements through a web-based control-room prototype. Its primary objective is to make autonomous agents more trustworthy, explainable, controllable, and accountable during operation.

The proposed interface follows a human-in-the-loop approach. Rather than allowing autonomous agents to operate without supervision, the control room provides visibility into agent status, trust scores, activities, self-corrections, human interventions, audit events, supporting evidence, operating principles, and Safe Mode controls.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

The main dashboard, referred to as the Trust Posture dash-board, provides a high-level representation of the current trust and safety state of the system. The prototype displays a Trust Score of 94.6 based on 411 observed actions, with three of four agents active, 18 self-corrections, and two interventions. These values are treated as demonstration outputs of the prototype rather than results from a statistically validated real-world deployment.

The major contributions of this work are:

- A control-room framework for monitoring trustworthy autonomous multi-agent systems.
- Integration of human-in-the-loop controls with autonomous agent execution.
- Trust monitoring using trust scores, confidence, activity, corrections, and interventions.
- Evidence-based decision-making and auditability mechanisms.
- A Safe Mode mechanism for controlling autonomous execution.
- An evaluation methodology based on purposive sampling and operational observations.
- Consideration of ethical issues and potential dataset bias.

II. System Overview and Architecture

The proposed system is implemented conceptually as a web-based control room for monitoring and governing autonomous agents. The architecture consists of six major layers:

1) Multi-Agent System
2) Monitoring Layer
3) Safety and Governance Layer
4) Decision Layer
5) Human Oversight
6) Audit and Learning Layer

The overall operational flow is:

Autonomous Agents → Agent Actions → Evidence and Confidence Evaluation → Trust Assessment → Policy/Guardrail Check → Decision → Approve/Review/Correct/Pause → Audit Trail → Trust Update → Adaptive Improvement

This continuous loop enables the system to evaluate agent behaviour while maintaining mechanisms for correction and human intervention.

A. Multi-Agent Layer

The multi-agent layer consists of specialized autonomous agents responsible for different tasks. The prototype demonstrates four agents.

- **Atlas / Research**: Evidence synthesis.
- **Mira / Triage**: Incident classification.
- **Wright / Release**: Change verification.
- **Coda / Customer**: Response drafting.

The specialization of agents allows different responsibilities to be monitored individually rather than treating the entire multi-agent system as a single autonomous entity.

B. Monitoring Layer

The monitoring layer collects information about agent activity, trust scores, confidence, run counts, status, corrections, interventions, and audit events. This information forms the basis for evaluating operational behaviour.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

C. Safety and Governance Layer

The safety and governance layer applies operational controls such as Safe Mode, review requirements, pause/resume controls, and policy or guardrail checks. These mechanisms prevent unrestricted autonomous execution.

D. Decision Layer

Following evidence and confidence evaluation, an agent action can be approved, reviewed, corrected, or paused. This pro-vides multiple decision states instead of treating autonomous actions as simply successful or unsuccessful.

E. Human Oversight Layer

Human operators can inspect agent information, intervene in decisions, pause or resume agents, and provide confirmation when Safe Mode is active.

Table 1. Prototype Trust Posture Metrics.

Metric	Prototype Value
Trust Score	94.6
Observed Actions	411
Active Agents	3 of 4
Self-Corrections	18
Human Interventions	2

Table 2. Agent Roster Demonstration Values.

Agent	Role	Status	Trust
Atlas	Research	Running	96.8
Mira	Triage	Running	93.4
Wright	Release	Review	88.1
Coda	Customer	Paused	91.7

F. Audit and Learning Layer

Meaningful decisions, corrections, approvals, reviews, and evidence updates are recorded through an audit trail. The resulting information can be used to update trust assessments and support adaptive improvement.

III. Trustworthy Multi-Agent System

A. Trust Posture Dashboard

The Trust Posture dashboard provides a high-level representation of system trust and operational safety. It displays the overall Trust Score, observed actions, active agents, self-corrections, interventions, trust trajectory, and adaptive corrections.

The prototype interface displays the following demonstration values:

The dashboard also includes a Trust Trajectory visualization for monitoring changes in trust relative to an operating threshold, as shown in Figure 1.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

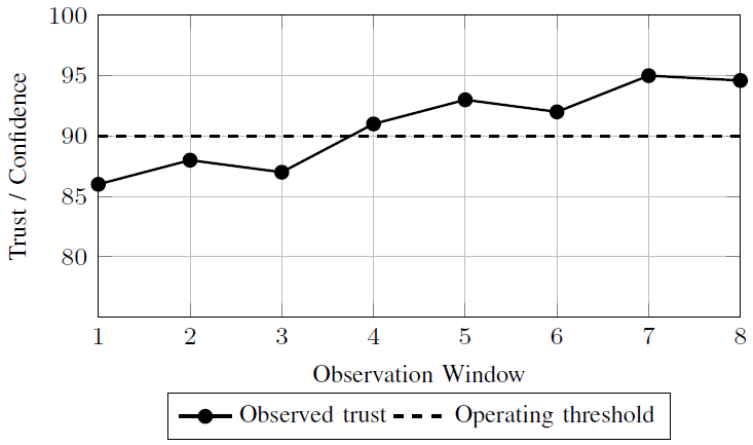


Figure 1. Illustrative trust trajectory used to demonstrate the proposed moni-toring interface. The plotted values are prototype visualization values and are not statistically validated experimental measurements.

B. Agent Roster

The Agent Roster provides individual-level information about each autonomous agent. It displays the agent's role, operational status, trust score, and execution count.

The prototype displays Atlas / Research with 184 runs, Mira/ Triage with 72 runs, Wright / Release with 39 runs, and Coda/ Customer with 116 runs. Operators can pause or resume agents, providing an explicit mechanism for human control.

The displayed trust scores are visualized in Figure 2.

Figure 2 illustrates the difference in displayed trust levels across the four prototype agents. Wright / Release has the lowest displayed trust score and is shown in Review status.

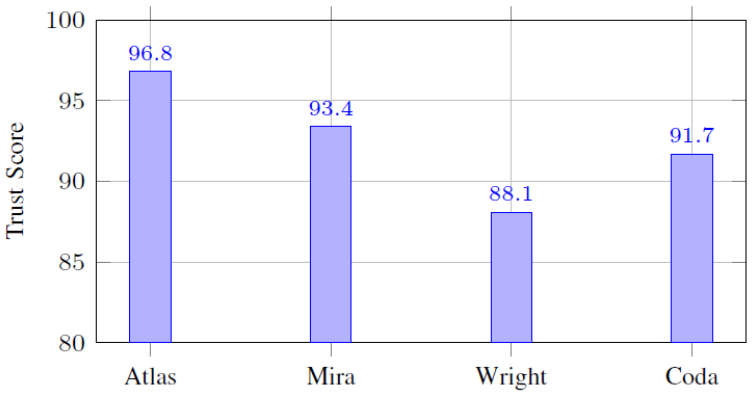


Figure 2. Agent-level trust scores displayed by the prototype. The values represent demonstration outputs rather than independently validated experimental measurements.

C. Audit Events and Operating Principles

The audit system maintains a human-readable record of meaningful decisions and corrections. Example events include action approval, self-correction, human review, and evidence refresh.

The system follows three principal concepts:

1) **Evidence Before Action**
2) **Confidence Is Not Certainty**
3) **Correction Is a Feature**

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

These principles emphasize that autonomous decisions should be supported by evidence, confidence should not be interpreted as absolute certainty, and correction should be incorporated into normal autonomous operation.

D. Safe Mode

Safe Mode is a central human-oversight mechanism. When Safe Mode is active, new autonomous actions require operator confirmation. This provides controlled autonomy while maintaining human responsibility for consequential actions.

The human-in-the-loop governance process is illustrated in Figure 3.

The operational behaviour is represented as:

Safe Mode ON → Autonomous Action → Human Confirmation → Action (1)

When Safe Mode is disabled:

Safe Mode OFF → Autonomous Action → Action May Proceed ATuhtoemsaetlieccatlelyd (2)

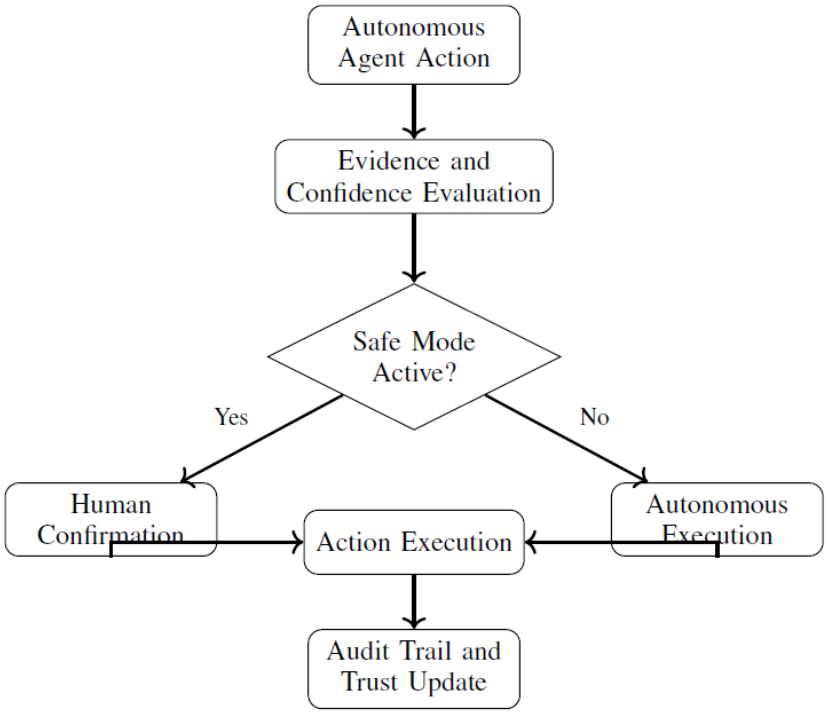


Figure 3. Human-in-the-loop governance flow of the proposed Trustworthy Multi-Agent System. Safe Mode determines whether autonomous actions require explicit human confirmation before execution.

IV. Measurement Concepts and Evaluation Metrics

Measurement is used to assign values to variables in order to evaluate research outcomes. In this work, measurement concepts are applied to assess trustworthiness, reliability, activity, and human oversight within the Trustworthy Multi-Agent System.

A. Independent Variable

The independent variable is the implementation of the Trustworthy Multi-Agent System, including:

- Trust monitoring
- Evidence-based decisions
- Adaptive corrections
- Auditability
- Safe Mode controls

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

B. Dependent Variables

The system can be evaluated using:

- **T**rust Score
- Confidence Level
- Agent Activity
- Self-Correction Count
- Human Intervention Count
- Agent Status
- Audit Events
- Decision/Action Status

These variables provide measurable indicators of autonomous-agent behaviour and operational trustworthiness.

C. Levels of Measurement

A metrics enable analysis of both autonomous activity and human involvement within the system.

Table 3. Levels of Measurement.

Level	Example
Nominal	Agent Role: Research, Triage, Release, Customer
Ordinal	Agent Status: Running, Review, Paused
Interval	Confidence / Trust trajectory values
Ratio	Runs, observed actions, self-corrections, interventions

V. Sampling Methodology

A. Sampling Technique

The proposed research uses **Purposive Sampling**. Under this approach, important agent activities and decision events are intentionally selected for evaluation. The objective is to focus on events that provide meaningful evidence about trust, correction, human intervention, and evidence-based decision-making.

B. Sampling Categories

The selected activities include:

- Agent actions
- Self-correction events
- Human interventions
- Approval events
- Review requests
- Evidence refresh events
- Paused or resumed activities
- Confidence-related events

C. Rationale for Purposive Sampling

Purposive sampling is appropriate because the research focuses on meaningful system events rather than randomly selecting activities. Trust and safety mechanisms are most clearly evaluated through events such as corrections, interventions, reviews, approvals, and evidence updates.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

VI. Dataset Analysis and Feature Identification

A. Dataset Description

The evaluation is designed around the MAESTRO bench-mark for multi-agent systems. The benchmark provides execution and system-level information that can be used to characterize multi-agent workloads and operational behaviour. The dataset used for quantitative analysis must be treated separately from the prototype-specific dashboard values. The trust scores, human-intervention counts, self-correction counts, and other dashboard values described in Section III are demonstration variables of the proposed control-room interface. They are therefore not treated as ground-truth labels from MAESTRO.

If the benchmark data are used for quantitative evaluation, the exact dataset version, files, preprocessing procedure, number of records, and selected variables should be reported before experimental statistics are presented.

Table 4. Proposed Evaluation Features.

Feature	Description
Agent Identifier	Identifies the participating agent
Agent Role	Defines the agent's responsibility when avail-able
Execution Record	Represents an observed execution or trace
System Metric	Represents an observed operational measurement
Resource Metric	Represents resource utilization when available
Event Information	Records relevant system events
Timestamp	Indicates temporal ordering of observations

B. Dataset Components

For the proposed evaluation, relevant information can be organized into:

1) Agent Activity Records
2) Execution and Trace Information
3) System Monitoring Measurements
4) Agent-Level Information
5) Resource-Utilization Information
6) Event Records

Only variables that are actually present in the downloaded benchmark should be included in the final quantitative analysis.

C. Feature Identification

These features can support analysis of agent behaviour and system operation. Prototype-only variables such as Trust Score, Self- Correction, and Human Intervention should not be presented as benchmark ground truth unless they are explicitly derived from validated labels or measurable events.

VII. Data Collection, Ethics, and Dataset Bias

A. Data Collection

Data collection for the quantitative component consists of extracting the required observations from the selected bench-mark files. The workflow should include dataset acquisition, data inspection, preprocessing, feature selection, aggregation where necessary, and evaluation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

The prototype control-room observations are retained as interface demonstrations and should be clearly distinguished from benchmark measurements.

B. Ethical Considerations

The proposed system incorporates responsible AI principles through:

- Human oversight for important autonomous actions.
- Transparency through monitoring and audit events.
- Accountability through recorded decisions and corrections.
- Controlled autonomy through Safe Mode.
- Evidence-based decisions before consequential actions.
- Explainability through information about decisions and system activity.

These mechanisms help ensure that autonomy does not eliminate human responsibility or control.

C. Dataset Bias

Dataset bias can occur when collected observations do not represent the full range of possible agent behaviours. Potential sources include:

- Limited numbers of observed activities.
- Unequal activity across agents.
- Greater representation of normal cases.
- Limited examples of failures.
- Limited human-intervention events.
- Benchmark-specific workload characteristics.

D. Bias Mitigation

Potential mitigation strategies include:

1) Including observations from multiple agents.
2) Analysing different categories of system events.
3) Including both normal and exceptional situations when available.
4) Applying consistent preprocessing and evaluation procedures.
5) Reporting the distribution of observations across agents.
6) Clearly documenting the benchmark version and selected files.

VIII. Results and Discussion

The prototype demonstrates a continuous monitoring and governance loop in which autonomous agents perform tasks, their actions are evaluated using evidence and confidence information, their trust posture is monitored, and the system can approve, review, correct, pause, or request human intervention. The displayed Trust Score of 94.6 across 411 observed actions demonstrates how overall trust information can be represented within the proposed control-room interface. Similarly, the presence of 18 self-corrections and two interventions demonstrates how corrective behaviour and human oversight can be surfaced as explicit operational metrics. These values are prototype demonstration outputs and are not treated as independent experimental measurements.

At the agent level, different displayed trust scores and operational states demonstrate the importance of evaluating agents individually. For example, Wright / Release is shown in Review status with a trust score of 88.1, while Atlas / Research is shown as Running with a trust score of 96.8.

The Safe Mode mechanism further demonstrates controlled autonomy. Instead of completely removing autonomous ca-pabilities, the system introduces an intermediate governance mechanism

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

in which autonomous actions require human confirmation. This creates a balance between automation and human control.

The benchmark-based quantitative results should be re-ported only after the selected MAESTRO data have been processed and the corresponding evaluation methodology has been applied. No accuracy, precision, recall, F1-score, or confusion matrix is claimed in this version because a validated classification target and ground-truth labels have not yet been established.

Overall, the prototype demonstrates the integration of runtime monitoring, evidence-based decisions, adaptive self-correction, human oversight, and auditability within a unified interface.

IX. Limitations

The current work has several limitations.

First, the control-room observations are based on a prototype demonstration rather than a large-scale real-world deployment. Consequently, the displayed dashboard metrics can-not independently establish the effectiveness of the proposed framework.

Second, the benchmark-based evaluation depends on the exact MAESTRO files selected, the available variables, and the preprocessing procedure. These details must be reported explicitly for the experimental results to be reproducible.

Third, the available observations may have an uneven distribution across agents, workloads, or operating conditions. Such imbalance can affect comparative analysis.

Fourth, normal system behaviour may be more strongly rep-resented than failures, unusual events, or high-risk situations. This can limit the ability to evaluate the framework under challenging conditions.

Finally, the current framework describes trust-related measurements but does not establish a universal or standardized trust-scoring methodology. Prototype trust scores should there-fore not be interpreted as ground-truth measures of system reliability.

X. Future Scope

Future development can extend the prototype into a larger-scale evaluation framework for autonomous multi-agent systems.

Potential directions include:

- **Real-world deployment:** Evaluate the framework using operational multi-agent systems rather than demonstration data.
- **Larger datasets:** Collect substantially more agent actions, interventions, corrections, and failure cases.
- **Failure-oriented evaluation:** Increase representation of exceptional and unsafe scenarios.
- **Advanced trust modelling:** Develop comprehensive approaches for calculating and updating agent trust scores.
- **Automated anomaly detection:** Identify unusual agent behaviours and automatically trigger review mechanisms.
- **Historical analysis:** Maintain long-term trust trajectories for individual agents.
- **Expanded human oversight:** Provide configurable intervention policies based on action risk and confidence.
- **Improved auditability:** Connect decisions with evidence, confidence, corrections, and human approvals.
- **Comparative evaluation:** Compare autonomous operation with different levels of human oversight and Safe Mode configurations.

XI. Conclusion

This paper presented a web-based control-room prototype for a Trustworthy Multi-Agent System designed to make autonomous agents more transparent, controllable, explainable, and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 2 November 2026 https://doi.org/10.20944/preprints202611.0098.v1

accountable. The framework combines trust monitoring, evidence-based decision-making, adaptive self-correction, run-time monitoring, human oversight, auditability, and Safe Mode controls.

The system provides multiple levels of visibility through the Trust Posture dashboard, Agent Roster, Recent Audit Events, and Safe Mode confirmation mechanism. The pro-posed measurement framework considers trust, confidence, activity, self-corrections, interventions, status, audit events, and action status as measurable indicators of system behaviour. Purposive sampling is used to focus evaluation on meaningful events such as agent actions, corrections, interventions, approvals, reviews, and evidence updates. The proposed dataset structure captures agent identity, roles, execution information, system measurements, events, and resource-related observations.

The framework also incorporates responsible AI principles by maintaining human oversight, transparency, accountability, controlled autonomy, evidence-based decision-making, and explainability. At the same time, the study recognizes potential dataset bias and the limitations associated with prototype-based observations.

The overall framework demonstrates that trustworthy autonomous agents require more than autonomous task execution. They require continuous monitoring, evidence-aware decision-making, mechanisms for correction, human control, and auditable operational records. The proposed control-room architecture provides a foundation for developing and evaluating safer and more accountable multi-agent systems.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Reference

1. Verified scholarly references on trustworthy artificial intelligence, multi-agent systems, human-in-the-loop AI, explainable AI, runtime monitoring, AI governance, and the selected MAESTRO benchmark should be inserted here.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.