

# Benchmark Reporting

## Transparency & Validity

- R.1. Is fully or at least partially open-sourced.
- R.2. Offers an open-source evaluation harness for users.
- R.3. Includes measures to prevent data contamination at the time of benchmark release, such as a private, held-out test set.
- R.4. Includes measures or plans to consistently update challenges over time to avoid overfitting.
- R.5. Clearly states the relationship between the agent capabilities it aims to evaluate and the constructs or outcomes it measures.
- R.6. Clearly states the evaluation subjective of the benchmark (e.g., a model or an agent framework).

## Flaw Mitigation

- R.7. Describes steps taken to prevent, identify, and correct flaws.
- R.8. Includes qualitative discussions of the potential impact of unavoidable flaws.

## Interpretation

- R.9. Includes quantitative analysis to assess the impact of unavoidable flaws (e.g., noise of ground truth).
- R.10. Reports metrics about statistical significance, such as confidence intervals.
- R.11. Provides guidance on interpreting results with eval flaws.
- R.12. Reports results of non-AI baselines (e.g., human experts).
- R.13. Reports results of trivial agents (e.g., one that does nothing).