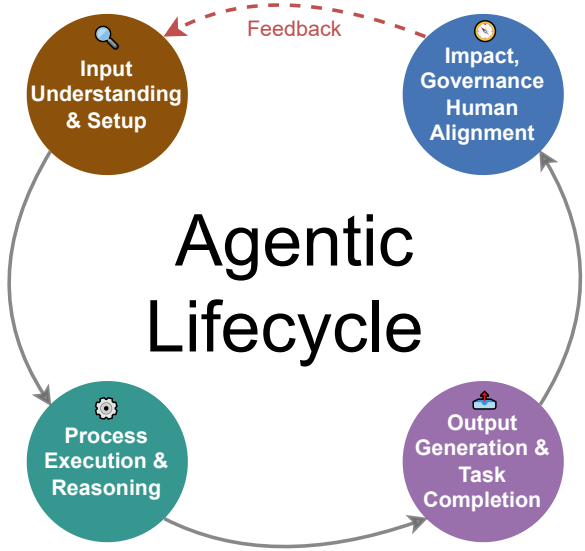

Ethics & Robustness

Fairness / Bias Mitigation · Cooperative Behavior · Adaptability · Robustness · Harm Reduction Score (when interpreted as post-impact risk control)

Transparency & Accountability

Explainability · Transparency · Rule Fidelity · Graph Edit Distance (GED) · Process Trace Audit (represented implicitly from ScaleAI ToolComp)

Retrieval & Grounding Quality

Context Relevance, Retrieval Precision, Average Precision, Hit Rate, Reciprocal Rank, NDCG, Prompt Safety Risk, Topic Relevance

Tool & Policy Initialization

Tool-Call Syntactic Accuracy, Parameter Accuracy, Tool Selection Accuracy, Policy Adherence (Stateful Evaluation / τ-Bench), Completion Under Policy (CuP), Risk Ratio (Policy Violations), Safety Metrics (HAP, PII, Bias, Harm, etc.) · Unsuccessful Requests

User Experience & Utility

User Satisfaction / Net Promoter Score · Click-Through Rate (CTR) · Gross Merchandise Value (GMV) · Readability (Text Ease / Grade Level) · Context Faithfulness / Answer Similarity (from WatsonX)

Accuracy & Graded Performance

Accuracy · Precision · Recall · F1 Score · Must Include · Must Exclude · eval_vqa · eval Fuzzy Image Match · Answer F1 (AgentBench) · Multi-hop F1 / Supporting Fact F1 (HotPotQA)

Planning & Action Correctness

Action F1 · Step Success Rate · Plan Success Rate · Partial Correctness · Progress (Lateral Puzzles / ALFWorld Progress Rate) · Adaptive Multi-Dimensional Score · Goal Drift · Tool Execution Success · Turn-Wise Advantage

Success & Completion Metrics

Success Rate (MLAgentBench, AgentBench, MINT, etc.) · Task Success Rate · Exact Match · Quasi Exact Match (GAIA) · Weighted Accuracy (FieldWorkArena) · Partial Completion (Mind2Web2) · Landmark Metric (SUPER) · Fuzzy Match

Reasoning Reliability & Supervision

Process Supervision Score · LLM Grading · Agent-as-Judge Score · Script-Executed Metric · Joint Accuracy (TRAIL) · Error Detection Rate (TRAIL) · Resource-Scaling Metrics · Pass@k / Pass^k · Harm Reduction Score

Temporal & Resource Efficiency

Token Usage · Wall-Clock Time · Task Completion Time (TCT) · Average Improvement · Improvement Rate · Raw Competition Scores · Reward Score

Domain-Specific Outputs

Operation F1 / Element Accuracy (Mind2Web) · Attribute F1 (WebShop) · File-Level Localization / CST Node-Level Retrieval (SWE-PolyBench) · Resolved Rate (SWE-Bench) · Price Comparison Accuracy / Cross-Shop Success (WebMall) · Tool Execution Success (re-evaluated for output) · Pass@k (reused for final outcome) · Win Rate / Reward Score (competition tasks) · Progress Rate (ALFWorld) · Average Improvement / Improvement Rate (performance gain)