
SCALING AGENTIC CAPABILITIES, NOT CONTEXT: EFFICIENT REINFORCEMENT FINETUNING FOR LARGE TOOLSPACES

Karan Gupta

Pranav Vajreshwari

Yash Pandya

Akshay Nambi

Ahmed Awadallah

Microsoft Research

Corresponding author: akshayn@microsoft.com

ABSTRACT

Agentic systems operating over large tool ecosystems must plan and execute long-horizon workflows under weak or non-verifiable supervision. While frontier models mitigate these challenges through scale and large context budgets, small language models (SLMs) remain brittle: eager tool loading saturates context, execution errors compound over time, and sparse rewards limit learning. We introduce *ATLAS*, a reinforcement finetuning framework that enables SLMs to operate effectively in large-scale toolspace environments by learning *how* to acquire context and *how* to execute actions. Our approach makes two key contributions. First, we treat context control and execution structure as learnable decisions, combining iterative tool loading with programmatic tool orchestration to bound context growth and stabilize long-horizon trajectories. Second, we propose rubric-based reinforcement finetuning, which decomposes task success into structured, task-aligned criteria and enables scalable training using small judge models. Across MCP benchmarks, these design choices yield large and consistent gains over generic RL baselines, allowing a 4B SLM to approach frontier-agent performance under far tighter parameter and context budgets.

1 INTRODUCTION

Agentic systems are increasingly deployed in *Model Context Protocol (MCP)* environments, where a single request requires coordinated planning, reasoning, and execution across multiple tools and services Anthropic (2024). While MCP enables large-scale integration, it shifts the core challenge to **decision-making at scale**: selecting relevant tools, maintaining state over long trajectories, and operating under tight context and cost budgets Yao et al. (2022); Shinn et al. (2023). In practice, agents interact with hundreds of tools across servers, and exposing full tool registries upfront leads to reasoning over large, heterogeneous schemas, causing context bloat and degraded efficiency Jia & Li (2025); Cai et al. (2023); Wang et al. (2024a).

Frontier systems address this via *dynamic tool discovery* and *programmatic tool calling (PTC)*, retrieving tools on-demand and using executable orchestration to manage control flow while keeping large artifacts out of prompts Wu et al. (2025); Jones & Kelly) (2025); Schmid (2026). However, these approaches assume high context capacity and robust execution, which do not hold for *small language models (SLMs)* Lu et al. (2024); Hoffmann et al. (2022). As a result, directly adopting such designs exposes structural failure modes Kim et al. (2025), highlighting that the key limitation is not reasoning ability, but a mismatch between frontier-scale assumptions and the constraints of efficient models.

Challenges for SLMs in live MCP environments. Agentic systems in MCP face *tool-space explosion*, *long-horizon execution*, and *weak supervision*, which disproportionately impact SLMs Jia & Li (2025). Agents must reason over large, heterogeneous tool sets under tight context budgets, making tool discovery brittle and costly Cai et al. (2023). Long-horizon workflows amplify early

errors due to limited context for tracking goals and state Shinn et al. (2023), while code-based orchestration exposes weaknesses in synthesis and recovery Belcak et al. (2025). Moreover, MCP tasks lack a single verifiable outcome, making outcome-only rewards too sparse for effective credit assignment and leading to brittle behaviors Cai et al. (2025); Tang et al. (2025). These challenges indicate that the key limitation is not model capacity, but the lack of mechanisms to regulate *context growth and execution structure*, motivating a shift toward learning *what context to acquire, when to acquire it, and how to represent execution compactly* Yao et al. (2022).

We introduce *ATLAS (Adaptive Tool Loading and Scoped Context)*, a *reinforcement finetuning (RFT)* framework for enabling efficient agentic reasoning with SLMs in large-scale MCP environments. *ATLAS* is complementary to frontier agent architectures in that it adopts the same high-level principles of on-demand tool access and structured execution, while explicitly targeting the efficiency-constrained regime where context, computation, and supervision are scarce. Its core novelty lies in treating *context acquisition and execution structure as learnable decisions*, optimized through reinforcement learning rather than fixed architectural choices.

ATLAS is built around two abstractions. First, **adaptive tool loading** exposes a compact capability overview and incrementally materializes detailed tool schemas only when required, bounding the tool search space at each step while supporting multi-server workflows. Second, **orchestration through code** represents long-horizon tool use as executable programs rather than turn-by-turn natural language interactions, enabling compact state representation and reliable control flow without repeatedly surfacing intermediate results to the model. While these abstractions define how efficient agentic behavior can be represented, learning such behavior remains challenging due to non-verifiable objectives and delayed supervision. Existing RFT approaches for agents typically rely on frontier LLM judges producing synthetic trajectory-level rewards for dimensions such as task success or tool correctness Cai et al. (2025); Zhao et al. (2025); Guo et al. (2025); Singh et al. (2025b). Although effective at small scale, these rewards are coarse, obscure task-specific structure, and require repeated use of expensive frontier judges, limiting scalability.

ATLAS introduces **rubric-based reinforcement finetuning**, where an LLM judge evaluates agent behavior using structured, task-level rubrics instead of directly producing a single score. These rubrics decompose evaluation along task-aligned criteria such as completeness, grounding in tool outputs, appropriateness of tool choices, and parameter precision; providing a richer learning signal without step-level annotations. Crucially, once rubrics are generated, *evaluation can be reliably performed by SLMs*, enabling scalable RFT without continued reliance on frontier judges. We further introduce an automated pipeline for rubric generation, allowing task-specific supervision to be produced once and reused throughout training.

We evaluate *ATLAS* on a suite of synthetic but realistic MCP tasks constructed using live MCP servers Wang et al. (2025). The training set comprises approximately 300 tasks spanning 28 MCP servers, and the held-out evaluation set includes tasks from both seen and previously unseen servers. We find that task-level rubric rewards yield more stable and effective reinforcement finetuning than generic outcome rewards Singh et al. (2025b); Guo et al. (2025). Notably, under rubric-based evaluation, an open-source *SLM judge* (Qwen3-30B-Instruct) outperforms a GPT-4o-based generic judge and matches stronger frontier judges such as o4-mini. When combined with programmatic orchestration, *ATLAS* consistently outperforms JSON-style tool calling in both task completion and efficiency. Our key contributions are

1. We introduce *ATLAS*, a reinforcement finetuning framework that enables efficient agentic behavior through adaptive context control and code-based orchestration for long-horizon tool use.
2. We show that learning these behaviors is essential: task-level, structured rubric rewards generated by an LLM judge provide stable supervision under non-verifiable objectives without requiring step-level annotations.
3. We demonstrate that rubric-based evaluation enables SLMs to serve as effective judges, outperforming GPT-4o-based generic judging, enabling scalable and cost-efficient reinforcement finetuning.
4. We show that task-level rubrics consistently outperform generic outcome rewards, and that code-based orchestration outperforms JSON-style tool calling across diverse MCP tasks.

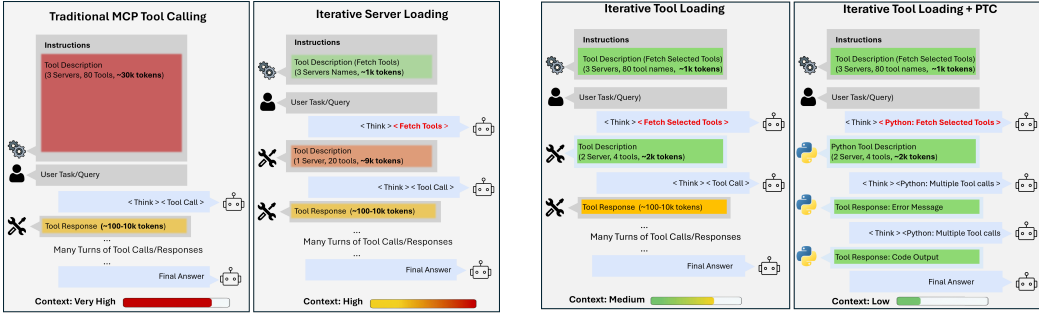


Figure 1: Context growth and execution structure across MCP agent designs. Traditional MCP agents incur high context costs by loading all tools upfront. ISL and ITL progressively reduce context by scoping server and tool schemas, while ITL+PTC further minimizes prompt growth by moving execution state into programmatic orchestration.

2 ATLAS OVERVIEW

We present ATLAS, a framework for training agentic systems to operate efficiently over large and evolving tool spaces under strict context and cost constraints. Rather than eagerly loading all tool schemas, ATLAS structures agent execution around an inference loop that incrementally materializes only the tool information required at each step.

ATLAS treats *context acquisition as a decision*: the agent explicitly selects which domains, servers, and tools to load based on the current task state, enabling long-horizon, multi-server workflows without excessive context growth. Execution is represented via *code-based orchestration*, expressing tool use and control flow as compact executable programs instead of turn-by-turn natural language, keeping intermediate state out of the prompt while preserving explicit control flow. These behaviors are learned end-to-end through reinforcement finetuning.

2.1 ITERATIVE SERVER LOADING (ISL)

ATLAS treats server selection as an explicit decision. At the start, the agent is given a compact index of available MCP servers and a meta-operation to retrieve tool schemas for a selected server. Conditioned on the task and current execution state, the agent selects a server, materializes only the tools exposed by that server, and executes the required tool calls, loading additional servers incrementally only as needed. This staged exposure avoids eager loading across all servers, preserves context for execution-critical information, and bounds tool selection to a single server at a time, enabling adaptive expansion of the tool space while maintaining a controlled context footprint.

2.2 ITERATIVE TOOL LOADING (ITL)

Iterative server loading enables reasoning over server-level structure but does not ensure grounding in the concrete capabilities of individual tools. In MCP settings, servers often expose hundreds of tools with verbose schemas, making eager materialization impractical and introducing substantial context overhead. ATLAS addresses this through **Iterative Tool Loading (ITL)**, which separates high-level planning from detailed tool grounding. Upon loading a server, the agent initially observes only a compact list of tool names, enabling lightweight reasoning about capabilities and plan structure without committing context to full schemas. As execution proceeds, the agent selectively materializes detailed schemas only for the tools required at a given decision point, incrementally grounding the plan as needed. By deferring full tool loading until use, ITL preserves context for execution-critical information while enabling scalable reasoning over large and heterogeneous tool collections.

2.3 UNIFIED PROGRAMMATIC TOOL ORCHESTRATION

Iterative server and tool loading regulate *what* context an agent sees, but conventional JSON-based tool calling still couples reasoning and execution through turn-by-turn natural language interaction. Control flow must be re-derived at each step, intermediate outputs are repeatedly injected into the context, and execution state is tracked only implicitly, leading to rapid context growth and brittle failure modes in long-horizon workflows.

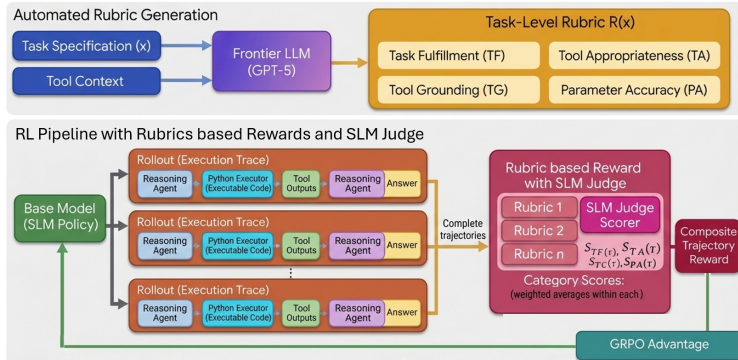


Figure 2: ATLAS Reinforcement Finetuning approach with Rubrics as Rewards and SLM Judge.

ATLAS replaces this with **unified programmatic execution model**, where all tool interactions are mediated by a persistent Python interpreter. Tool calls are expressed as function invocations, control flow is encoded explicitly using programming constructs, and intermediate results are stored in program state rather than surfaced to model, enabling compact, stateful execution without context overhead. Importantly, this is not equivalent to exposing Python as an additional tool. The interpreter serves as a unified orchestration layer across MCP servers, with execution proceeding via synthesis and refinement of a single program. Errors are handled through localized program edits rather than replaying reasoning traces. Combined with iterative tool loading, this yields disciplined execution in which schemas are materialized only when needed, state remains within the program, and only task-relevant outputs are exposed, which substantially improve robustness and scalability in long-horizon, multi-server settings.

2.4 SCAFFOLDING FOR PROGRAMMATIC TOOL ORCHESTRATION

MCP tool libraries are out of distribution relative to a model’s coding knowledge, and without additional structure, models often fail to translate plans into correct executable code. Errors such as incorrect function names, mismatched arguments, or improper use of tool outputs can derail long-horizon execution even when high-level reasoning is sound. ATLAS introduces a lightweight Python-side scaffolding layer that enables robust programmatic orchestration across previously unseen MCP servers. The scaffold standardizes tool representation, invocation, and debugging, providing a stable execution interface that bridges abstract reasoning and executable behavior.

The scaffold exposes a uniform `MCPServer` abstraction that presents MCP tools as Python-native functions. Tool schemas from heterogeneous servers are normalized into consistent function signatures, verified against underlying APIs, and bound dynamically, insulating agents from JSON-specific idiosyncrasies. Tool outputs are converted into Python-native data structures and retained in program state rather than re-injected into the model context; when output schemas are missing, representative structures are inferred and reused. To support Iterative Tool Loading, a `get_tools_info()` interface materializes full function signatures and examples only for selected tools. Even when loaded, these Python-native definitions are substantially more token-efficient than raw JSON, enabling precise grounding while preserving context. Additional implementation details and design choices for the scaffold are provided in Appendix A.

3 REINFORCEMENT FINETUNING

Reinforcement learning for agentic MCP tasks is challenging due to long-horizon, interdependent decision-making. Correct behavior depends on sequences of tool selection, argument construction, execution order, and grounded use of tool outputs, where early errors can invalidate downstream steps despite a plausible final answer. These challenges are compounded by *non-verifiable objectives*: MCP tasks rarely have a single ground truth, and multiple execution trajectories may satisfy user intent, making supervision based on final answers or binary success sparse and often misleading. Effective learning therefore requires reward signals that capture procedural quality and task-specific constraints rather than outcomes alone.

Limitations of LLM-as-Judge. Most LLM-as-judge frameworks assign a single score, often accompanied by brief reasoning or critique, to an entire agent trajectory Wang et al. (2025). For complex,

long-horizon tasks with multiple interacting requirements, this coarse evaluation is insufficient: successful task fulfillment typically depends on satisfying several distinct behavioral and outcome-level criteria. This limitation is particularly acute when LLM-as-judge evaluations are used as rewards in GRPO-style training Guo et al. (2025), where group-relative advantage estimation assumes consistent and comparable scoring across trajectories for the same task. In practice, variability in the judge’s internal reasoning and strictness can lead to inconsistent relative rankings, injecting noise into the learning signal and undermining stable policy optimization.

3.1 TASK-LEVEL RUBRICS FOR AGENTIC RFT

Agentic MCP tasks are long-horizon and require satisfying multiple interdependent functional criteria. In this regime, generic trajectory-level judgments are difficult for LLM judges to apply consistently, leading to unstable supervision under long contexts and stochastic generation.

We therefore replace direct LLM judge with with LLM judge under the guidance of **task-level structured rubrics**. For each task, we create a set of rubrics that explicitly define success for each coarse-grained criteria, including task fulfillment, tool appropriateness, grounding in tool outputs, and parameter correctness, together with a weight for each criterion.

At evaluation time, these fixed rubrics are provided to the LLM judge alongside each trajectory, which assigns scores per rubric. The weighted combination of these scores yields the final reward used for training. By holding rubrics constant across all trajectories for a given task, this approach enforces consistent relative judgments, reduces variance in reward signals, and improves supervision quality compared to unconstrained trajectory-level evaluation.

Automated Rubric Generation. Manual rubric design is not scalable in MCP settings due to task diversity and heterogeneous tool usage. ATLAS therefore generates rubrics automatically *once per task* using a frontier LLM (GPT-5) offline. A rubric is defined as a set of weighted criteria: $\mathcal{R}(x) = \{(C_i, D_i, W_i)\}_{i=1}^N$, where x denotes the task specification and C_i, D_i and W_i correspond to criteria name, criteria description and criteria weightage respectively. The rubric generation prompt (see Appendix F.1) only includes the task specification and available tool context, it is independent of any particular agent trajectory, avoiding bias introduced by long or noisy execution traces.

Each criteria folds under one of four categories capturing core requirements of agentic tool use: task fulfillment, tool appropriateness, grounding in tool outputs, and parameter accuracy. To reduce ambiguity and variance, rubric generation is constrained so that criteria are (i) observable from execution traces, (ii) non-overlapping, and (iii) aligned with functional task requirements rather than surface-level language quality.

RFT with Rubric-Based Rewards. We adopt a cold-start reinforcement finetuning setup in which all supervision is provided by rubric-based rewards, without step-level annotations or supervised warm-start. For each task x , a task-specific rubric $\mathcal{R}(x)$ is generated offline using GPT-5 and fixed throughout training. All rollouts for the same task are evaluated against the same rubric, enabling stable relative comparisons during advantage estimation under GRPO.

Each rubric criteria is scored, and then it’s weighted sums are normalized under the category it folds into. There are four categories for trajectory quality: (i) **Task Fulfillment (TF)**; (ii) **Tool Appropriateness (TA)**; (iii) **Tool Grounding (TG)**; (iv) **Parameter Accuracy (PA)**.

For a trajectory τ , an LLM-based judge assigns a score $d_i(\tau) \in [0, 1]$ to each rubric criterion C_i , with 1 indicating full satisfaction. Each rubric criterion C_i is associated with a weight W_i . For each category $R \in \{\text{TF, TA, TG, PA}\}$ containing N_R criteria, we compute a weighted category score: $S_R(\tau) = \frac{\sum_{i=1}^{N_R} W_i d_i(\tau)}{\sum_{i=1}^{N_R} W_i}$. The final trajectory reward is a *composite reward* obtained by aggregating category scores: $R(\tau) = \sum_{R \in \{\text{TF, TA, TG, PA}\}} \alpha_R S_R(\tau)$, with fixed category weights α_R shared across all rollouts for the task, assigning higher weight to task fulfillment. Examples of the automated rubrics generation are provided in Appendix E.

Scaling Judging from Frontier LLMs to SLMs. A key advantage of rubric-based rewards is scalability. In prior LLM-as-judge approaches, evaluating long trajectories with frontier models often dominates training cost. Rubric-conditioned scoring reduces judging from holistic trajectory

assessment to evaluating concrete, well-specified criteria, substantially lowering the reasoning burden on the judge.

As a result, small language models (SLMs) can serve as effective judges when guided by task-level rubrics. We find that SLM judges provide stronger and more stable learning signals than frontier LLM judges using generic rewards, and achieve better performance than stronger frontier judges (e.g., GPT4o) under rubric-based evaluation. This enables scalable and cost-efficient reinforcement finetuning for long-horizon, non-verifiable agentic MCP tasks.

During optimization, gradients are masked over tool-output tokens so that learning applies only to model-generated content. This focuses policy updates on planning, tool selection, argument construction, execution control, and termination decisions, rather than imitation of tool responses.

4 EXPERIMENTAL SETUP

4.1 MCP TASKS AND ENVIRONMENTS

We evaluate *ATLAS* on synthetic but realistic agentic tasks over live MCP servers, requiring long-horizon reasoning, tool discovery and invocation, and grounding in tool outputs. Tasks are non-verifiable, admitting multiple valid execution trajectories.

Train data. Our task data generation is derived from MCPBench Wang et al. (2025), from which we select 28 MCP servers spanning heterogeneous tool ecosystems (search, structured data, file systems, and enterprise APIs). Using MCPBench’s synthetic task proposal pipeline, we generate over 1,000 multi-server tasks with 1–3 servers per task and stratified difficulty (Easy, Medium, Hard), corresponding to approximately 2–3, 4–5, and 6+ distinct requirements. Tasks are generated using `o4-mini` and filtered with automated judges to retain high-quality tasks (solvability > 9 , utility > 8). Since many generated tasks are either unsolvable with the available tools or exceed the practical limits of SLMs, we further filter tasks by executing them with Kimi K2 Thinking (1T parameters) and retain only those achieving a task fulfillment score of at least 4/10, yielding a final set of 304 solvable, high-quality tasks (see Table 6 in Appendix D.2).

Test data. We evaluate generalization on two held-out test sets. The first consists of 104 unseen tasks from the MCPBench benchmark drawn from previously seen 28 servers, measuring task-level generalization. The second, *ATLAS-Test*, comprises 100 tasks drawn from 11 unseen MCP servers + 28 MCP servers, evaluating server-level and tool-level generalization under distribution shift. These tasks are designed to assess generalization to unseen servers and novel tool combinations. Appendix D presents the details of various servers used in MCPBench and *ATLAS-Test* along with task distribution.

4.2 MODELS

We evaluate *ATLAS* on open-weight language models spanning different parameter scales and levels of agentic capability, (i) **Qwen2.5-7B-Instruct** Yang et al. (2024) (7B), chosen for strong instruction-following and suitability for RL and (ii) **Qwen3-4B-Instruct-2507** Yang et al. (2025) (4B), a smaller model with native tool-calling support with 32K context length for both models.

As a strong frontier baseline, we additionally evaluate **Kimi K2 Thinking** Team et al. (2025) (1T parameters, 80K Context), configured with traditional MCP execution in which all available tool schemas are eagerly loaded into the system prompt. We further evaluate Kimi under our execution scaffolding with **Iterative Server Loading (ISL)** and **Iterative Tool Loading (ITL)** to isolate the impact of context control and execution structure independent of model scale.

4.3 EVALUATION METRICS

We evaluate models performance on MCP tasks using an **LLM-as-judge** framework. We adopt the evaluation prompt provided by MCPBench Wang et al. (2025), which scores complete agent trajectories given the task specification and requirements. The judge evaluates performance across four categories: **Task Fulfillment (TF)**—whether core task requirements are satisfied; **Tool Appropriateness (TA)**—whether selected tools are relevant and necessary; **Tool Grounding (TG)**—whether tool outputs are used faithfully and correctly; and **Parameter Accuracy (PA)**—the correctness and

# Variant	MCPBench Task Ful.	Avg Turns	Avg Tokens
Kimi-K2 Thinking (Frontier Baseline)			
1 All Tools Loaded	4.38	20	23768
2 Iterative Server Loading	4.11	27	21747
3 Iterative Tool Loading	3.62	28	18290
4 PTC	5.62	33	33796
Qwen2.5-7B-Instruct (Server Loading)			
4 Iterative Server Loading	2.33	25	11192
5 + RL w/ GPT-4o Judge (Generic)	3.02	19	10307
6 + RL w/ Qwen3-30B Judge (Generic)	2.84	24	12201
7 + RL w/ GPT-4o Judge (Rubrics)	3.13	19	11882
8 + RL w/ Qwen3-30B Judge (Rubrics)	3.18	21	13075
Qwen3-4B-Instruct (Server Loading)			
9 Iterative Server Loading	2.73	24	9152
10 + RL w/ GPT-4o Judge (Generic)	3.25	17	9372
11 + RL w/ Qwen3-30B Judge (Generic)	3.12	21	11530
12 + RL w/ GPT-4o Judge (Rubrics)	3.43	18	10654
13 + RL w/ Qwen3-30B Judge (Rubrics)	3.87	19	11142
Qwen3-4B-Instruct (Tool Loading)			
14 Iterative Tool Loading	2.36	20	9045
15 + PTC	2.94	18	13462
16 + RL w/ Qwen3-30B J. (Generic)	3.48	23	12815
17 + PTC + RL w/ Qwen3-30B J. (Generic)	3.91	20	12951
18 + RL w/ Qwen3-30B J. (Rubrics)	4.03	20	11151
19 + PTC + RL w/ Qwen3-30B J. (Rubrics)	4.15	18	13400
Qwen3-30B-Instruct			
20 Iterative Server Loading	2.51	28	9849
21 + RL w/ Qwen3-30B Judge (Rubrics)	4.08	21	9575
22 Iterative Tool Loading	2.62	44	15334
23 + RL w/ Qwen3-30B Judge (Rubrics)	4.19	24	12936
24 Iterative Tool Loading + PTC	3.11	18	16234
25 + RL w/ Qwen3-30B Judge (Rubrics)	4.44	21	12201

Table 1: Performance comparison across model families, execution strategies, and reinforcement finetuning configurations. **Gray**: base (no learning); **Blue**: frontier judge (GPT-4o); **Green**: SLM judge (Qwen3-30B); **Orange**: Programmatic Tool Calling (PTC). Task Fulfillment (TF) is reported on a 0–10 scale.

precision of tool arguments. Following MCPBench, we use `o4-mini` as the LLM judge for its robustness in validation. The full evaluation prompt is provided in the Appendix F.3.

4.4 REINFORCEMENT FINETUNING SETUP

Task-level rubrics are generated once per task using a frontier model (GPT-5) Singh et al. (2025a), while trajectory evaluation during training is performed by an **SLM judge** (Qwen3-30B-Instruct) Qwen Team (2025) conditioned on these rubrics. Generic LLM Judge reward baselines are evaluated using frontier judges (GPT-4o) Hurst et al. (2024).

We compare agent variants differing only in execution and learning mechanisms: (i) traditional MCP execution with eager tool loading, (ii) iterative server loading (ISL), (iii) iterative server and tool loading (ISL+ITL), and (iv) ATLAS (ISL+ITL with programmatic orchestration).

Implementation Details. All experiments are conducted on machines equipped with $8 \times$ NVIDIA B200 GPUs. We use the `verl` library as the reinforcement learning framework and extend it to support MCP tool calling, including a custom rollout structure, truncation of overlong tool responses, and integration of LLM-based judge evaluation into the training loop. Reward computation and policy updates are performed online during training. All models share fixed hyperparameters and identical optimization settings to ensure fair comparison. Additional implementation details are provided in Appendix B.

5 RESULTS

5.1 OVERALL PERFORMANCE ON MCPBENCH

Table 1 reports **Task Fulfillment (TF)** on MCPBench across model families, execution strategies, and reinforcement finetuning configurations. TF (0–10) is our primary metric and directly measures

end-to-end success on compositional, long-horizon MCP tasks. Additional diagnostic metrics and ATLAS-Test results are reported in Appendix C, Table 4 and 5.

Key takeaway. Structured execution alone is insufficient for small models, but when combined with reinforcement finetuning—especially rubric-based supervision—efficient SLMs approach frontier-level performance.

Across all model scales, base agents without learning perform poorly, typically achieving TF in the 2–3 range (Qwen2.5-7B ISL: 2.33, Row 4; Qwen3-4B ISL: 2.73, Row 9; Qwen3-4B ITL: 2.36, Row 14). Introducing reinforcement finetuning yields large and consistent gains, while structured execution amplifies these improvements by stabilizing long-horizon trajectories.

In contrast, the frontier model **Kimi-K2 Thinking** achieves strong performance even with all tools eagerly loaded (TF 4.38, Row 1), reflecting its ability to tolerate large context windows. However, it does *not* benefit from ITL (TF drops to 3.62, Row 3), as it is not trained to reason over restricted, dynamically loaded tool views. Importantly, **PTC significantly improves frontier performance**, increasing TF to 5.62 (Row 4), demonstrating that structured, programmatic execution provides gains even for large models.

5.2 EFFECT OF EXECUTION STRUCTURE

We first isolate the impact of execution structure under cold-start conditions, holding learning fixed.

Iterative Server Loading (ISL). Under ISL, both SLMs exhibit low TF (Qwen2.5-7B: 2.33, Row 4; Qwen3-4B: 2.73, Row 9), confirming that server-level context control alone is insufficient for non-verifiable MCP tasks.

Iterative Tool Loading (ITL). Moving from ISL to ITL reduces schema exposure and token usage (Qwen3-4B: 9,152 → 9,045 tokens, Rows 9 → 14), but yields only modest gains in TF (2.73 → 2.36). This mirrors the behavior seen in Kimi (Rows 2–3): models not explicitly trained to reason over JSON-limited tool views struggle to fully exploit ITL under cold-start conditions.

Programmatic Tool Calling (PTC). Adding PTC under ITL improves execution robustness even without learning. For Qwen3-4B, TF increases from 2.36 to 2.94 (Rows 14 → 15), despite higher token usage due to executable code. This indicates that programmatic control flow reduces execution failures and stabilizes long-horizon behavior, providing a stronger substrate for learning.

Key takeaway. Without learning, ITL yields modest gains (~10%), while PTC delivers stronger improvements (~25%) by stabilizing execution—showing that structure alone is insufficient.

5.3 IMPACT OF REINFORCEMENT FINETUNING

We now examine the effect of reinforcement finetuning (RFT), holding execution structure fixed.

ISL + RFT. Under ISL, RFT produces large gains across both model scales. For Qwen2.5-7B, TF improves from 2.33 (Row 4) to 3.02 with a generic frontier judge (Row 5), and further to 3.18 with rubric-based supervision and an SLM judge (Row 8). Qwen3-4B exhibits the same pattern, improving from 2.73 (Row 9) to 3.25 (Row 10) and up to 3.87 with rubric-based RFT (Row 13).

ITL + RFT. Under ITL, generic RFT substantially improves performance (Qwen3-4B: 2.36 → 3.48, Rows 14 → 16). However, JSON-based ITL remains less effective than its PTC counterpart, even after learning.

ITL + PTC + RFT (Best-performing SLM regime). The strongest results are obtained when ITL, PTC, and RFT are combined. For Qwen3-4B, TF reaches 3.91 under generic RFT with an SLM judge (Row 17), outperforming JSON-based ITL variants. Applying rubric-based rewards with PTC further improves performance to 4.15 (Row 19), closely matching frontier model Kimi performance (Rows 1–3). We also observe **low variance across repeated evaluations** (std ≈ 0.0575) with close agreement to reported scores, indicating stable and reliable gains.

Key takeaway. RFT delivers the largest gains (+35–65%), with ITL+PTC providing the strongest execution substrate for learning.

5.4 RUBRIC-BASED REWARDS AND JUDGE SCALING

Rubrics vs. generic rewards. Across models and execution strategies, **rubric-based RFT consistently outperforms generic scalar rewards**. Under ISL, Qwen2.5-7B improves from **3.02** to **3.18** (Rows 5 → 8), while Qwen3-4B shows larger gains from **3.25** to **3.87** (Rows 10 → 13). Similar improvements persist under tool loading: Qwen3-4B improves from **3.48** to **4.03** without PTC, and from **3.91** to **4.15** with PTC (Rows 16–19). These results demonstrate that **task-aligned, structured rewards are critical for effective credit assignment in non-verifiable MCP tasks**.

Consistency of rubric generation. A separate source of variability arises from rubric generation, which impacts training quality rather than evaluation. To assess robustness, we generate multiple rubric sets using the same rubric generator (GPT-5) and observe consistent downstream performance (row 13): TF scores of **3.87**, **3.91**, and **3.79**, with a mean of **3.86** and low variance (std **0.05**). This demonstrates that **rubric-based training is stable with respect to the specific rubric instantiation**, provided the generator model remains fixed.

Open-source rubric generation. We further evaluate rubric generation using strong open-source models (row 13). While GPT-5 achieves the best performance (**3.85**), open-source alternatives remain competitive: Kimi-K2 obtains **3.57** and Qwen3.5-27B achieves **3.47**. These results show that **rubric-based training does not depend on proprietary models**, and that open-source generators can produce effective, though slightly lower-quality, training signals.

Frontier vs. SLM judges. Under generic rewards, GPT-4o and Qwen3-30B perform comparably (e.g., Qwen2.5-7B: Rows 5 vs. 6). Under rubric-based supervision, however, the **SLM judge consistently outperforms the frontier judge** (Qwen3-4B: **3.87** vs. **3.43**, Rows 13 vs. 12). Since all methods are evaluated using a fixed external judge (o4-mini), these gains reflect improved training dynamics rather than evaluation bias.

Key takeaway. Rubric-based supervision improves task success by up to **20%** and allows SLM judges to outperform frontier judges.

Interaction B/W Execution Structure and Learning A central finding is the strong **positive interaction** between execution structure and learning. Structured execution without learning yields limited gains (Rows 14–15), while RFT under weak structure saturates early (Rows 5–8, 10–13). In contrast, combining disciplined execution (ITL+PTC) with RFT produces the highest TF, indicating that execution structure enables effective credit assignment over long horizons rather than acting as a standalone improvement. We also observe that performance improves with increased training data, with gains saturating beyond moderate dataset sizes (see Appendix C).

Efficiency and Training Dynamics ISL and ITL significantly reduce token usage by constraining schema exposure (Qwen3-4B: **12,906** tokens under ISL, Row 9, vs. **9,045** under ITL, Row 14). PTC increases token usage due to executable representations (Row 15), but reduces interaction turns and enables higher TF once learning is applied (Rows 16–17). The added cost reflects explicit execution structure and is offset by improved task success.

Scaling to larger models. We further evaluate the proposed recipe on a larger model, **Qwen3-30B-Instruct**. Across execution strategies, **rubric-based RFT consistently yields substantial gains**. Under iterative server loading, TF improves from **2.51** to **4.08** (Rows 20→21), while under iterative tool loading it increases from **2.62** to **4.19** (Rows 22→23). Combining tool loading with PTC further boosts performance from **3.11** to **4.44** (Rows 24→25), achieving the best overall result. These results show that the proposed framework **scales effectively with model size**.

5.5 CLOSING THE GAP TO FRONTIER MODELS

The strongest SLM configuration—**Qwen3-4B with ITL, PTC, and RFT (Rubrics)**—achieves **4.15 TF** (Row 19), closely approaching frontier **Kimi-K2 Thinking** under eager loading (**4.38**, Row 1) and matching its performance under structured settings such as ISL (**4.11**, Row 2). This is achieved despite operating with **orders of magnitude fewer parameters** and **far tighter context budgets**. However, we still lag behind Kimi under **PTC** (**5.62**, Row 4), where strong coding priors in frontier models provide an advantage.

Key takeaway. With structured execution and rubric-based RFT, a 4B SLM matches or approaches frontier performance under ITL/ISL/eager settings, highlighting that **training and structure can close much of the gap to frontier agents**, though advantages remain under PTC due to stronger coding capabilities.

6 RELATED WORK

Large language models have increasingly evolved from static generators into interactive agents capable of multi-step reasoning, tool use, and environment interaction. ATLAS is most closely related to work on learned tool invocation, reinforcement learning for agentic behavior, adaptive context and tool selection, and programmatic execution.

Learned Tool Invocation. Early work established mechanisms for augmenting LLMs with external tools. Toolformer demonstrated self-supervised API call insertion to improve downstream performance Schick et al. (2023). More recent approaches treat tool calls as executable code, enabling stateful, programmatic interaction and stronger control over long-horizon tasks Wang et al. (2024b).

Reinforcement Learning for Agentic Tool Use. Reinforcement learning has been widely used to teach agents when and how to invoke tools. ReTool applies outcome-based RL to interleave reasoning with real-time code execution without step-level supervision Feng et al. (2025a). ARTIST unifies reasoning, tool selection, and environment interaction within an RL framework for autonomous tool use Singh et al. (2025b). Related work explores learning stability and credit assignment, including turn-level rewards Zeng et al. (2025) and modeling interactive user dynamics Zhao et al. (2025).

Dynamic Tool Selection and Context Control. As static tool inventories become impractical, several approaches focus on adaptive tool selection. AutoTool trains models to dynamically select tools using large-scale rationale data and dual-phase optimization Jia & Li (2025). ToolSample combines curriculum learning with dynamic sampling Feng et al. (2025b), while Tool-Star coordinates multiple tools under RL supervision Dong et al. (2025), highlighting the importance of structured training for scalable tool use. MCP-Bench evaluates tool-using agents in MCP environments, exposing challenges from multi-server tool ecosystems and large context footprints Wang et al. (2025).

Programmatic Execution. Programmatic orchestration has emerged as an alternative to turn-by-turn natural language tool calls. Executable action frameworks reduce context overhead and enable stateful, verifiable control flow Wang et al. (2024b). This paradigm underlies systems such as Anthropic’s programmatic tool calling interfaces Anthropic (2025); Wu et al. (2025), MCP tooling like `mcp-cli` Schmid (2026), and terminal-based agent environments such as SETA CAMEL-AI (2026).

Evaluation and Supervision. Sparse rewards remain a major challenge for RL in complex agentic settings. Recent work shows that structured, rubric-based evaluation provides denser supervision and improves training stability in non-verifiable domains Yu et al. (2025).

ATLAS. Prior work addresses individual aspects of agentic reasoning—tool invocation, reinforcement learning, dynamic selection, or programmatic execution. ATLAS differs by jointly treating *context acquisition as a learned decision* and *execution structure as a first-class optimization target* for efficiency-constrained agents operating over large MCP tool spaces. By combining adaptive tool loading, compact programmatic orchestration, and task-level rubric-based reinforcement finetuning, ATLAS provides an integrated approach not explored in prior work.

7 CONCLUSION

This work argues that the central challenge in agentic MCP systems is not model scale, but how agents acquire context, execute actions, and learn from weak supervision. ATLAS addresses this by learning context control and execution structure through adaptive tool loading, programmatic orchestration, and task-level rubric supervision. Empirically, execution structure alone improves efficiency but yields limited task success; substantial gains arise only when reinforcement finetuning provides task-aligned credit assignment. Structured rubrics are central to this effect, enabling small judge models to match or surpass frontier judges and making training scalable and cost-efficient. As a result, a 4B SLM recovers nearly 90% of frontier-agent performance despite operating under far tighter parameter and context budgets. These results mark a shift from scale-driven to structure- and learning-driven agent design.

REFERENCES

- Anthropic. Introducing the model context protocol. <https://www.anthropic.com/news/model-context-protocol>, Nov 25 2024. Accessed on January 29, 2026.
- Anthropic. Programmatic tool calling (ptc) — claude developer docs. Documentation, 2025. URL <https://platform.claude.com/docs/en/agents-and-tools/tool-use/programmatic-tool-calling>. Programmatic tool calling allows Claude to write code that calls tools programmatically to reduce context usage.
- Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.
- Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *arXiv preprint*, arXiv:2510.00915, 2025. URL <https://www.arxiv.org/pdf/2510.00915>. arXiv:2510.00915 [cs.LG].
- CAMEL-AI. Seta: Scaling environments for terminal agents. Blog post, 2026. URL <https://www.camel-ai.org/blogs/seta-scaling-environments-for-terminal-agents>. Discusses asynchronous agents and environment scaling for terminal agent workflows.
- Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoners via reinforcement learning. *arXiv preprint arXiv:2505.16410*, 2025. URL <https://arxiv.org/abs/2505.16410>.
- Jiazhan Feng, Shijue Huang, Ge Zhang, Yujia Qin, Baoquan Zhong, et al. Retool: Reinforcement learning for strategic tool use in language models. *arXiv preprint arXiv:2504.11536*, 2025a. URL <https://arxiv.org/abs/2504.11536>.
- Zihao Feng, Xiaoxue Wang, Bowen Wu, Hailong Cao, Tiejun Zhao, Qun Yu, and Baoxun Wang. Toolsample: Dual dynamic sampling methods with curriculum learning for rl-based tool learning. *arXiv preprint arXiv:2509.14718*, 2025b. URL <https://arxiv.org/abs/2509.14718>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jingyi Jia and Qinbin Li. Autotool: Efficient tool selection for large language model agents, 2025. URL <https://arxiv.org/abs/2511.14650>.
- Anthropic (Adam Jones and Conor Kelly). Code execution with mcp: Building more efficient agents. <https://www.anthropic.com/engineering/code-execution-with-mcp>, Nov 2025. Accessed on January 29, 2026.
- Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Mark Malhotra, et al. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025.

-
- Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.
- Qwen Team. Qwen3-30B-Instruct. <https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507>, 2025. Hugging Face model card.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. URL <https://arxiv.org/abs/2302.04761>.
- Philipp Schmid. mcp-cli: Lightweight cli to interact with mcp servers. GitHub repository, 2026. URL <https://github.com/philschmid/mcp-cli>. A lightweight command-line interface for discovering and interacting with MCP servers.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025a.
- Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*, 2025b.
- Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. Beyond verifiable rewards: Scaling reinforcement learning for language models to unverifiable data. *arXiv preprint arXiv:2503.19618*, 2025.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.
- Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*, 2024a.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*, 2024b. URL <https://arxiv.org/abs/2402.01030>.
- Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, Yujia Bao, and Eugene Siow. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. *arXiv preprint arXiv:2508.20453*, 2025. URL <https://arxiv.org/abs/2508.20453>.
- Anthropic (Bin Wu, Adam Jones, Artur Renault, Henry Tay, Jake Noble, Nathan McCandlish, Noah Picard, and Sam Jiang). Introducing advanced tool use on the claude developer platform. Anthropic engineering blog, 2025. URL <https://www.anthropic.com/engineering/advanced-tool-use>. Feature announcement including programmatic tool calling and dynamic tool search.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, et al. Qwen2.5 technical report. *arXiv preprint, arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>. Includes instruction-tuned Qwen2.5 models within the Qwen2.5 family of LLMs.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin

-
- Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint*, arXiv:2505.09388, 2025. URL <https://arxiv.org/abs/2505.09388>. arXiv:2505.09388 [cs.CL].
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- Zhaochen Yu, Ling Yang, Jiaru Zou, Shuicheng Yan, and Mengdi Wang. Demystifying reinforcement learning in agentic reasoning. *arXiv preprint arXiv:2510.11701*, 2025. URL <https://arxiv.org/abs/2510.11701>.
- Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*, 2025. URL <https://arxiv.org/abs/2505.11821>.
- Weikang Zhao, Xili Wang, Chengdi Ma, Lingbin Kong, Zhaohua Yang, Mingxiang Tuo, Xiaowei Shi, Yitao Zhai, and Xunliang Cai. Mua-rl: Multi-turn user-interacting agent reinforcement learning for agentic tool use, 2025. URL <https://arxiv.org/abs/2508.18669>.

APPENDIX

A PROGRAMMATIC TOOL CALLING IMPLEMENTATION

A key piece of the puzzle to make PTC work is coming up with a scaffold atop the python executor that makes it possible for the SLM to go from an initial plan or pseudocode to a working code given that our MCPServer library and its tools are out-of-distribution for the model's coding knowledge.

While trying to write working code from the pseudocode/plan using unseen MCP Server function formats, the model can make three following types of mistakes:

1. Incorrect function names
2. Incorrect argument names and formats
3. Incorrect output access and composition.

To address these issues, we implement 1. Input schema normalization, 2. MCPServer class abstraction and dynamic function allocation, 3. Server output conversion 4. Function output schema and examples fetching 5. Iterative Tool Loading Implementation and 6. Informative error logs and hints.

A.1 INPUT SCHEMA NORMALIZATION

Different MCP servers tend to reveal different levels of information about their tools and may even use different JSON keys to describe similar properties. We run an LLM script to map these tool input schemas into a common, clean schema without losing any information. We then process these JSON schemas into Python function and argument names. This process is run in parallel for all tools and servers, and the results are stored offline. The fetched function names and argument names are programmatically verified to match the original JSON schemas.

Example python converted schema:

```
home_manager_search(query: str (required), limit: int (optional,
default=20, max=100))
```

Note that most MCP servers do not define the output schema, and at this stage, we also do not have the Python output formats.

A.2 MCPSEVER CLASS ABSTRACTION AND DYNAMIC FUNCTION ALLOCATION

The MCPServer class interfaces between python native code and the MCP servers' json API. The class is instantiated at a server level (e.g. `time_mcp = MCPServer("Time MCP")`). This class dynamically sets the attributes of the class based on the `server_name` its initialized with (e.g. `time_mcp.get_current_time(timezone='America/NewYork')`) and then maps them to the MCP server calls. For a specific server, all the tool names and the args are fetched from the saved mapping created in the previous step.

A.3 SERVER OUTPUT CONVERSION

The responses from the MCP servers always come as serialized text; we need to load them into appropriate python types so that they can be used for sequential processing in code. We use `ast.literal_eval()` for this conversion which follows a "top-down" approach, moving from structural complexity (lists/tuples/dicts) to primitive data types (int/float/str).

A.4 FUNCTION OUTPUT SCHEMA AND EXAMPLES FETCHING

Once we have the python function calling wrapper around the MCP servers' API, we can get the output schemas. We design a simple tool calling loop with an LLM endpoint and prompt it to execute the python functions with realistic arguments given the descriptions. The LLM calls the tools till it succeeds in getting the outputs, then it stores the output format and an example for each tool in the same python native format.

A.5 ITERATIVE TOOL LOADING IMPLEMENTATION

To enable orchestration across tool calls and servers, we need to give the model a good understanding of the exact tool usage. To achieve this, we implement a `mcp_server_obj.get_tools_info([])` function that fetches the full tool schemas and examples for the requested tool names from the pre-saved info in the last step. Note that even with the full schemas and examples, python definitions take a lot less tokens than raw json schemas.

A.6 INFORMATIVE ERROR LOGS AND HINTS

The logic for most agentic tasks is simple enough for the model to get. The challenge is in writing the code using a completely unseen library without taking too many trial-and-error steps. We nudge the model towards correcting itself in the following ways:

- **Incorrect function names:** Instead of a plain "object of class MCPServer has no attribute 'incorrect_func_name'", we give the message "MCP Server 'server_name' doesn't have the tool 'incorrect_func_name' available tools: 'list of tool names', did you mean closest tool?"
- **Argument names and formats:** In this case, most of the servers themselves give a clear feedback on the arg errors, we log the same for every failed function call in the code block.
- **Incorrect output access:** If the model encounters an error while using the returned value from any server as a variable, we pass its output schema in the logs. We also provide hints like "You have tried to access a string as a dict, please the output logs or use `get_tools_details[]`" instead of a plain "str indexes must be integers".

A.7 EXAMPLE ISL VS ITL+PTC BEHAVIOR

The Qwen3-4B ISL agent with json tool calling examples shows multiple tool calls and verbose intermediate reasoning when the steps have dependencies between them:

ISL Example 1

```
[user]
Task: I'm pulling together a report on last quarter's harvest from our 10 farms, and honestly I need some hard numbers. We recorded yields of 120, 150, 150, 200, 180, 170, 160, 140, 130, and 155 tons.
```

```
Here's what I'm trying to nail down: - What's our total output, average yield per farm, the median and the most common harvest size, plus our lowest and highest yields and the overall spread?
- Then, at $30 a ton, what does that translate to in revenue?
- After covering $2,000 in fixed costs per farm (so 10 farms total), what's left as net profit and what's our profit margin when you express it as a percentage (rounded to the nearest whole number)?
- Finally, I'm curious about the gap between our top-performing farm (200 tons) and the average yield—if that difference is more than 30 tons, I want to budget extra fertilizer at $10 per ton of that gap (and round up); if it's 30 or less, I'll stick with a $500 allowance (and round down).
```

```
Could you crunch all those figures? I really need solid data—can't go to my boss with just guesses. Thanks!
```

```
Please ensure all findings are supported by concrete data and verifiable sources. I need specific numbers and evidence, not generalizations..
```

```
The following servers are available:
Time MCP
Math MCP
```

```
[assistant]
```

```
<think>
First, I need to determine which server can provide the necessary calculations.
The Math MCP server seems to be the most relevant for this task as it can handle statistical and numerical operations.
</think>
```

```
<tool_call>
{"name": "fetch_tools", "arguments": {"server": "Math MCP"}}
</tool_call>
```

```
[user]
<tool_response>
Tools in server='Math MCP'
```

```
Tool: `Math MCP:add` (Server: Math MCP)
Description: Adds two numbers together
Input Schema:
```

```
{
  "type": "object",
  "properties": {
    "firstNumber": {
      "type": "number",
      "description": "The first addend"
    },
    "secondNumber": {
      "type": "number",
      "description": "The second addend"
    }
  },
  "required": [
    "firstNumber",
    "secondNumber"
  ],
  "additionalProperties": false
}
```

```
.....
</tool_response>
```

[assistant]

```
<tool_call>
{"name": "Math MCP:sum", "arguments": {"numbers": [120, 150, 150, 200,
180, 170, 160, 140, 130, 155]}}
</tool_call>
```

[user]

```
<tool_response>
1555
</tool_response>
```

...

[assistant]

```
<tool_call>
{"name": "Math MCP:median", "arguments": {"numbers": [120, 150, 150, 200,
180, 170, 160, 140, 130, 155]}}
</tool_call>
```

[user]

```
<tool_response>
undefined
</tool_response>
```

[assistant] It seems the median function did not return a value, which might be due to the way the function is implemented or the specific version of the tool. Let's calculate the median manually by sorting the list and finding the middle value.

First, let's sort the list of yields: [120, 130, 140, 150, 150, 155, 160, 170, 180, 200]. Since there are 10 values, the median will be the average of the 5th and 6th values.

Let's calculate the median.

```
<tool_call>
{"name": "Math MCP:mean", "arguments": {"numbers": [150, 155]}}
</tool_call>
```

[user]

```
<tool_response>
152.5
</tool_response>
```

[assistant]

```
<tool_call>
{"name": "Math MCP:mode", "arguments": {"numbers": [120, 150, 150,
200, 180, 170, 160, 140, 130, 155]}}
</tool_call>
```

[user]

```
<tool_response>
Entries (150) appeared 2 times
</tool_response>
```

[assistant]

<think >To summarize:

- Total output: 1555 tons
- Average yield per farm: 155.5 tons

```

- Median yield: 152.5 tons
- Most common yield (mode): 150 tons
- Minimum yield: 120 tons
- Maximum yield: 200 tons
- Total revenue: $46,650
- Net profit: $26,650
- Profit margin: 57%
- Budget for extra fertilizer: $450
</think >

<answer>
- Total output: 1555 tons
- Average yield per farm: 155.5 tons
- Median yield: 152.5 tons
- Most common yield (mode): 150 tons
- Minimum yield: 120 tons
- Maximum yield: 200 tons
- Total revenue: $46,650
- Net profit: $26,650
- Profit margin: 57%
- Budget for extra fertilizer: $450
</answer>

```

ISL Example 2

[user]

Task: So, here's the deal: I'm putting together a quick "breakout radar" for BTC-USDT, ETH-USDT, and ADA-USDT, and I really need hard numbers to back any call. What I'm wondering is:

- What's the current price vs. its average over roughly the past day?
- How far off is that in percentage terms?
- In the last 15 minutes, does it look like the coin's on an upswing or heading down?
- And over the last 5 minutes, has volume shot up or tanked compared to its recent average?
- Finally—based on all that—are any of these really cracking out into a breakout right now?

Could you pull the live data, run those calculations, and give me a concise summary (JSON, table, whatever) for each pair? I can't walk into my team meeting with gut feels—I need real, data-driven answers. Thanks!

Please ensure all findings are supported by concrete data and verifiable sources. I need specific numbers and evidence, not generalizations..

The following servers are available:

Time MCP
OKX Exchange

[assistant]

<think>

First, I need to determine if the tools available on the OKX Exchange server can provide the necessary live data for BTC-USDT, ETH-USDT, and ADA-USDT.

</think>

<tool_call>

```
{
  "name": "fetch_tools",
  "arguments": {
    "server": "OKX Exchange"
  }
}
```

</tool_call>

[user]

<tool_response>

Tools in server='OKX Exchange'

Tool: `OKX Exchange:get_price` (Server: OKX Exchange)
Description: Get latest price for an OKX instrument
Input Schema:

```
{
  "type": "object",
  "properties": {
    "instrument": {
      "type": "string",
      "description": "Instrument ID (e.g. BTC-USDT)"
    }
  },
  "required": [
    "instrument"
  ]
}
```

...
</tool_response>

[assistant]

```
<tool_call>
{"name": "OKX Exchange:get_price", "arguments": {"instrument": "BTC-USDT"}}
</tool_call>
```

[user]

```
<tool_response>
{
  "instrument": "BTC-USDT",
  "lastPrice": "87910.2",
  "bid": "87914.9",
  "ask": "87915",
  "high24h": "89086.9",
  "low24h": "86090",
  "volume24h": "5521.45498049",
  "timestamp": "2026-01-26T04:35:11.908Z"
}
</tool_response>
```

[assistant]

Based on the price data for BTC-USDT, we have the following information:

- Last Price: 87910.2
- Bid: 87914.9
- Ask: 87915
- High (24h): 89086.9
- Low (24h): 86090
- Volume (24h): 5521.45498049

Next, I will fetch the candlestick data for the last 15 minutes to determine if the coin is on an upswing or heading down.

```
<tool_call>
{"name": "OKX Exchange:get_candlesticks",
"arguments": {"instrument": "BTC-USDT", "bar": "15m", "limit": 1}}
</tool_call>
```

[user]

```
<tool_response>
[
  {
    "timestamp": "2026-01-26T04:30:00.000Z",
    "open": "87848.7",
    "high": "87943.3",
    "low": "87848.7",
    "close": "87914.9",
    "volume": "30.93625999",
    "volumeCurrency": "2719377.337798413"
  }
]
</tool_response>
```

[assistant]

Based on the candlestick data for the last 24 hours for BTC-USDT, we can calculate the average price.

From the data, the close prices for the last 24 hours are as follows (for simplicity, I'll use a few data points, but in practice, we would use all 1440 data points):

- 2026-01-26T04:35:00.000Z: 87894.7
- 2026-01-26T04:34:00.000Z: 87918.9
- 2026-01-26T04:33:00.000Z: 87875.8
- 2026-01-26T04:32:00.000Z: 87901
- 2026-01-26T04:31:00.000Z: 87903.7
- 2026-01-26T04:30:00.000Z: 87921.3
- 2026-01-26T04:29:00.000Z: 87848.7
- 2026-01-26T04:28:00.000Z: 87847.9
- 2026-01-26T04:27:00.000Z: 87759
- 2026-01-26T04:26:00.000Z: 87759
- 2026-01-26T04:25:00.000Z: 87767.5
- 2026-01-26T04:24:00.000Z: 87767.5
- 2026-01-26T04:23:00.000Z: 87617
- 2026-01-26T04:22:00.000Z: 87647.1
- 2026-01-26T04:21:00.000Z: 87647.9
- 2026-01-26T04:20:00.000Z: 87617.1
- 2026-01-26T04:19:00.000Z: 87582.4
- 2026-01-26T04:18:00.000Z: 87508.3
- 2026-01-26T04:17:00.000Z: 87490.1
- 2026-01-26T04:16:00.000Z: 87496.2
- 2026-01-26T04:15:00.000Z: 87490
- 2026-01-26T04:14:00.000Z: 87496
- 2026-01-26T04:13:00.000Z: 87490

```

- 2026-01-26T04:12:00.000Z: 87496
- 2026-01-26T04:11:00.000Z: 87490
- 2026-01-26T04:10:00.000Z: 87496
- 2026-01-26T04:09:00.000Z: 87490
- 2026-01-26T04:08:00.000Z: 87496
- 2026-01-26T04:07:00.000Z: 87490
- 2026-01-26T04:06:00.000Z: 87496
- 2026-01-26T04:05:00.000Z: 87490
- 2026-01-26T04:04:00.000Z: 87496
- 2026-01-26T04:03:00.000Z: 87490
- 2026-01-26T04:02:00.000Z: 87496
- 2026-01-26T04:01:00.000Z: 87490
- 2026-01-26T04:00:00.000Z: 87496

```

The average price over the past 24 hours is calculated as follows:

$$\text{Average Price} = \frac{2637000}{30} = 87900$$

The current price is 87910.2, and the average price over the past 24 hours is 87900. The percentage difference is:

$$\left(\frac{87910.2 - 87900}{87900} \right) \times 100 \approx 0.0116\%$$

Now, let's summarize the findings for BTC-USDT:

- Current Price: 87910.2
- Average Price (24h): 87900
- Percentage Difference: 0.0116%
- Trend in the last 15 minutes: Upswing
- Volume in the last 5 minutes: Low compared to recent average

```

<answer>
{
  "BTC-USDT": {
    "current_price": 87910.2,
    "average_price_24h": 87900,
    "percentage_difference": 0.0116,
    "trend_15m": "Upswing",
    "volume_5m": "Low compared to recent average"
  }
}
</answer>

```

For the same tasks, the Qwen3-4B ITL+PTC agent orchestrates tool calls within a single code-block by capturing intermediate reasoning in code itself:

PTC Example 1

[User] I'm pulling together a report on last quarter's harvest from our 10 farms, and honestly I need some hard numbers. We recorded yields of 120, 150, 150, 200, 180, 170, 160, 140, 130, and 155 tons.

Here's what I'm trying to nail down: - What's our total output, average yield per farm, the median and the most common harvest size, plus our lowest and highest yields and the overall spread? - Then, at \$30 a ton, what does that translate to in revenue? - After covering \$2,000 in fixed costs per farm (so 10 farms total), what's left as net profit and what's our profit margin when you express it as a percentage (rounded to the nearest whole number)? - Finally, I'm curious about the gap between our top-performing farm (200 tons) and the average yield—if that difference is more than 30 tons, I want to budget extra fertilizer at \$10 per ton of that gap (and round up); if it's 30 or less, I'll stick with a \$500 allowance (and round down).

Could you crunch all those figures? I really need solid data—can't go to my boss with just guesses. Thanks!

Please ensure all findings are supported by concrete data and verifiable sources. I need specific numbers and evidence, not generalizations.

The following servers are available that can be initiated as a MCPServer Object:

Server: 'Time MCP' Available Tools: get_current_time, convert_time

Server: 'Math MCP' Available Tools: add, subtract, multiply, division, sum, mean, median, mode, min, max, floor, ceiling, round

[Assistant] Plan

Given yields:

```
[120, 150, 150, 200, 180, 170, 160, 140, 130, 155]
```

I will:

1. Compute total output using sum.
2. Compute average yield using mean.
3. Compute median yield using median.
4. Compute mode (most common yield) using mode.
5. Find minimum and maximum yields using min and max.
6. Compute spread (range) as max - min.
7. Compute revenue at \$30 per ton: revenue = total_output * 30.
8. Compute total fixed costs: 10 farms * \$2000 = \$20,000.
9. Compute net profit: revenue - fixed_costs.
10. Compute profit margin: (net_profit / revenue) * 100, rounded to the nearest whole number.

11. Compute the gap between the top farm (200 tons) and the average yield.
12. Apply fertilizer budget rule:
 - If $\text{gap} > 30 \rightarrow \text{fertilizer cost} = \$10 \times \text{ceil}(\text{gap})$
 - If $\text{gap} \leq 30 \rightarrow \text{fertilizer cost} = \text{floor}(500)$

[Assistant] Code

```
<tool_call>
{"name": "code_interpreter", "arguments": {"code": "from MCPBench import MCPServer

# Initialize the Math MCP server
math_mcp = MCPServer('Math MCP')

# Given yield data
yields = [120, 150, 150, 200, 180, 170, 160, 140, 130, 155]

# Compute required metrics
total_output = math_mcp.sum(numbers=yields)
average_yield = math_mcp.mean(numbers=yields)
median_yield = math_mcp.median(numbers=yields)
mode_yield = math_mcp.mode(numbers=yields)
min_yield = math_mcp.min(numbers=yields)
max_yield = math_mcp.max(numbers=yields)
spread = max_yield - min_yield

# Revenue at $30 per ton
revenue_per_ton = 30
revenue = total_output * revenue_per_ton

# Fixed costs: $2000 per farm x 10 farms
fixed_costs = 10 * 2000

# Net profit
net_profit = revenue - fixed_costs

# Profit margin (as percentage)
profit_margin = math_mcp.round(number=(net_profit / revenue) * 100)

# Gap between top farm (200 tons) and average yield
gap = 200 - average_yield

# Fertilizer budget
if gap > 30:
    fertilizer_cost = math_mcp.ceiling(number=gap * 10)
else:
    fertilizer_cost = math_mcp.floor(number=500)

# Compile all results
results = {
    'Total Output': total_output,
    'Average Yield': average_yield,
    'Median Yield': median_yield,
    'Most Common Yield (Mode)': mode_yield,
    'Lowest Yield': min_yield,
    'Highest Yield': max_yield,
    'Spread': spread,
    'Revenue': revenue,
    'Fixed Costs': fixed_costs,
    'Net Profit': net_profit,
    'Profit Margin (%)': profit_margin,
    'Gap (Top - Average)': gap,
    'Fertilizer Budget': fertilizer_cost
}

results"}}}
```

PTC Example 2

[User] So, here's the deal: I'm putting together a quick "breakout radar" for BTC-USDT, ETH-USDT, and ADA-USDT, and I really need hard numbers to back any call. What I'm wondering is:

– What's the current price vs. its average over roughly the past day? – How far off is that in percentage terms? – In the last 15 minutes, does it look like the coin's on an upswing or heading down? – And over the last 5 minutes, has volume shot up or tanked compared to its recent average? – Finally—based on all that—are any of these really cracking out into a breakout right now?

Could you pull the live data, run those calculations, and give me a concise summary (JSON, table, whatever) for each pair? I can't walk into my team meeting with gut feels—I need real, data-driven answers. Thanks!

Please ensure all findings are supported by concrete data and verifiable sources. I need specific numbers and evidence, not generalizations.

Task: I'm putting together a quick "breakout radar" for BTC-USDT, ETH-USDT, and ADA-USDT, are any of these really cracking out into a breakout right now? Could you pull the live data, run those calculations, and give me a concise summary (JSON, table, whatever) for each pair?

The following servers are available that can be initiated as a MCPServer Object:

Server: 'Time MCP' Available Tools: get_current_time, convert_time

Server: 'OKX Exchange' Available Tools: get_price, get_candlesticks

[Assistant] Plan

1. Fetch current prices for BTC-USDT, ETH-USDT, and ADA-USDT using `get_price`.
2. Fetch candlestick data for the past day (`bar='1D'`) to compute the daily average price and assess price deviation.
3. Calculate percentage deviation of the current price from the daily average.
4. Determine breakout status based on price movement and volume trends.

[Assistant] Code

```
<tool_call>
{ "name ": "code_interpreter ", "arguments ": { "code ":
"from MCPBench import MCPServer

# Initialize the OKX Exchange server
okx_mcp = MCPServer("OKXExchange")

# Fetch current prices for BTC-USDT, ETH-USDT, ADA-USDT
prices = {
    "BTC-USDT": okx_mcp.get_price(instrument="BTC-USDT"),
    "ETH-USDT": okx_mcp.get_price(instrument="ETH-USDT"),
    "ADA-USDT": okx_mcp.get_price(instrument="ADA-USDT")
}

# Fetch 1-day candlestick data (for daily average)
daily_candlesticks = {
    "BTC-USDT": okx_mcp.get_candlesticks(instrument="BTC-USDT", bar="1D", limit=100),
    "ETH-USDT": okx_mcp.get_candlesticks(instrument="ETH-USDT", bar="1D", limit=100),
    "ADA-USDT": okx_mcp.get_candlesticks(instrument="ADA-USDT", bar="1D", limit=100)
}

# Extract and process data
price_data = {}
for instrument, price in prices.items():
    price_data[instrument] = {
        "current_price": float(price["lastPrice"]),
        "bid": float(price["bid"]),
        "ask": float(price["ask"]),
        "high24h": float(price["high24h"]),
        "low24h": float(price["low24h"]),
        "volume24h": float(price["volume24h"])
    }

# Calculate daily average price from 1D candlesticks
daily_avg_price = {}
for instrument, candles in daily_candlesticks.items():
    closes = [float(c["close"]) for c in candles]
    daily_avg_price[instrument] = sum(closes) / len(closes)

# Calculate percentage deviation from daily average
price_deviation = {}
for instrument in price_data:
    current = price_data[instrument]["current_price"]
    avg = daily_avg_price[instrument]
    price_deviation[instrument] = ((current - avg) / avg) * 100

# Combine all data into a final summary
breakout_radar = {}
for instrument in price_data:
    breakout_radar[instrument] = {
        "current_price": price_data[instrument]["current_price"],
        "daily_avg_price": daily_avg_price[instrument],
        "price_deviation_percent": price_deviation[instrument],
        "breakout_status": "breakout"
        if abs(price_deviation[instrument]) > 2
        else "no breakout"
    }

breakout_radar "}}
</tool_call>
```

```

# Initialize the OKX Exchange server
okx_mcp = MCPServer('OKX Exchange')

# Fetch current prices for BTC-USDT, ETH-USDT, ADA-USDT
prices = {
    'BTC-USDT': okx_mcp.get_price(instrument='BTC-USDT'),
    'ETH-USDT': okx_mcp.get_price(instrument='ETH-USDT'),
    'ADA-USDT': okx_mcp.get_price(instrument='ADA-USDT')
}

# Fetch 1-day candlestick data (for daily average)
daily_candlesticks = {
    'BTC-USDT': okx_mcp.get_candlesticks(instrument='BTC-USDT', bar='1D', limit=100),
    'ETH-USDT': okx_mcp.get_candlesticks(instrument='ETH-USDT', bar='1D', limit=100),
    'ADA-USDT': okx_mcp.get_candlesticks(instrument='ADA-USDT', bar='1D', limit=100)
}

# Fetch 15-minute candlestick data (for trend analysis)
recent_15m = {
    'BTC-USDT': okx_mcp.get_candlesticks(instrument='BTC-USDT', bar='1m', limit=15),
    'ETH-USDT': okx_mcp.get_candlesticks(instrument='ETH-USDT', bar='1m', limit=15),
    'ADA-USDT': okx_mcp.get_candlesticks(instrument='ADA-USDT', bar='1m', limit=15)
}

# Fetch 5-minute candlestick data (for volume analysis)
recent_5m = {
    'BTC-USDT': okx_mcp.get_candlesticks(instrument='BTC-USDT', bar='5m', limit=5),
    'ETH-USDT': okx_mcp.get_candlesticks(instrument='ETH-USDT', bar='5m', limit=5),
    'ADA-USDT': okx_mcp.get_candlesticks(instrument='ADA-USDT', bar='5m', limit=5)
}

# Extract and process data
price_data = {}
for instrument, price in prices.items():
    price_data[instrument] = {
        'current_price': float(price['lastPrice']),
        'bid': float(price['bid']),
        'ask': float(price['ask']),
        'high24h': float(price['high24h']),
        'low24h': float(price['low24h']),
        'volume24h': float(price['volume24h'])
    }

# Calculate daily average price from 1D candlesticks
daily_avg_price = {}
for instrument, candles in daily_candlesticks.items():
    closes = [float(c['close']) for c in candles]
    daily_avg_price[instrument] = sum(closes) / len(closes)

# Calculate percentage deviation from daily average
price_deviation = {}
for instrument in price_data:
    current = price_data[instrument]['current_price']
    avg = daily_avg_price[instrument]
    deviation = ((current - avg) / avg) * 100
    price_deviation[instrument] = deviation

# Analyze 15-minute trend (last close vs. first close)
recent_15m_trend = {}
for instrument, candles in recent_15m.items():
    if len(candles) >= 2:
        first_close = float(candles[0]['close'])
        last_close = float(candles[-1]['close'])
        trend = 'up' if last_close > first_close else 'down'
        recent_15m_trend[instrument] = trend
    else:
        recent_15m_trend[instrument] = 'insufficient data'

# Analyze 5-minute volume (last volume vs. recent average)
recent_5m_volume = {}
for instrument, candles in recent_5m.items():

```

Hyperparameter	Value
Train batch size	16
PPO mini-batch size	4
Max context window	31000
Rollout samples (n)	4
Rollout temperature	1.0
Advantage estimator	GRPO
Learning rate	1×10^{-6}
Optimizer	AdamW
LR schedule	Flat
Precision	bfloat16
Max tool calls	20
Max tool response length	4000
KL loss coefficient	0.001

Table 2: Training hyperparameters.

```

if len(candles) >= 1:
    last_volume = float(candles[0]['volume'])
    # Estimate recent average volume from last 5 candles (if available)
    volumes = [float(c['volume']) for c in candles]
    avg_volume = sum(volumes) / len(volumes)
    volume_change = ((last_volume - avg_volume) / avg_volume) * 100
    recent_5m_volume[instrument] = volume_change
else:
    recent_5m_volume[instrument] = 'insufficient data'

# Combine all data into a final summary
breakout_radar = {}
for instrument in price_data:
    breakout_radar[instrument] = {
        'current_price': price_data[instrument]['current_price'],
        'daily_avg_price': daily_avg_price[instrument],
        'price_deviation_percent': price_deviation[instrument],
        '15m_trend': recent_15m_trend[instrument],
        '5m_volume_change_percent': recent_5m_volume[instrument],
        'breakout_status': 'breakout' if (abs(price_deviation[instrument]) > 2
            and recent_15m_trend[instrument] == 'up'
            and recent_5m_volume[instrument] > 0) else 'no breakout'
    }

breakout_radar "}}

</tool_call>

```

B TRAINING HYPERPARAMETERS

Table 2 summarizes the training and hyperparameters used throughout all our experiments.

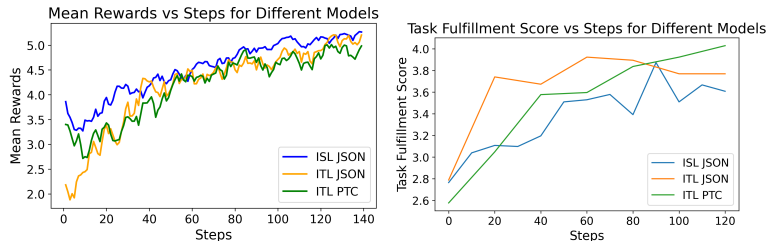


Figure 3: ATLAS reinforcement finetuning training and validation curves under ISL, ITL, and ITL+PTC, showing composite training rewards (left) and task-fulfillment-based validation performance (right).

C DETAILED RESULTS

Table 4 reports Task Fulfillment (TF), Grouping on Tool Results, Tool Appropriateness, Parameter Accuracy, Avg Turns, and Avg Tokens on MCPBench across model families, execution strategies, and reinforcement finetuning configurations.

Table 5 reports Task Fulfillment (TF), Grouping on Tool Results, Tool Appropriateness, Parameter Accuracy, Avg Turns, and Avg Tokens on the heldout ATLAS dataset across model families, execution strategies, and reinforcement finetuning configurations.

Training and Validation Graphs. Figure 3 shows that PTC and ITL achieve slightly lower training rewards across steps, yet outperform other variants on the target task fulfillment (TF) metric at test time. This divergence indicates stronger generalization to unseen tasks and suggests that these methods produce trajectories that are less susceptible to reward hacking the training-time judge, prioritizing true task success over optimizing the train reward.

Scaling with training data. We evaluate the effect of increasing training set size on Qwen3-4B with ISL and rubric-based RFT. As shown in Rows 26–28, performance improves from **3.48** at 150 examples to **3.87** at 306 examples, indicating that additional data significantly enhances learning. Increasing the dataset further to 606 examples yields a marginal improvement to **3.88**, suggesting **early saturation**. These results indicate that rubric-based RFT is **data-efficient**, achieving most gains with a relatively modest amount of training data.

Table 3: Effect of training data size on Qwen3-4B with ISL + Rubrics.

Train Size	Task Ful. \uparrow	Avg Turns \downarrow	Avg Tokens \downarrow
150	3.48	25	10953
306	3.87	11	11412
606	3.88	24	12351

Model	Variant	Task Fulfillment	Grounding	Tool Appropriateness	Parameter Accuracy	Avg Turns	Avg Tokens
Kimi-K2 Thinking	All Tools Loaded	4.38	4.50	4.99	4.82	20	23767
	Iterative Server Loading	4.11	4.40	4.84	4.72	27	21746
	Iterative Tool Loading	3.62	3.96	4.16	3.46	28	18290
Qwen2.5-7B-Instruct	Iterative Server Loading	2.33	3.019	3.82	4.10	25	11192
	+ RL w/ GPT-4o Judge	3.02	3.61	4.59	5.07	19	10306
	+ RL w/ Qwen3-30B Judge	2.84	4.61	5.10	5.30	23	12200
	+ RL w/ GPT-4o Rubrics	3.13	3.78	4.61	5.53	19	11882
	+ RL w/ Qwen3-30B Rubrics	3.18	3.38	4.34	5.11	21	13075
Qwen3-4B-Instruct	Iterative Server Loading	2.73	4.49	4.35	5.06	24	9152
	+ RL w/ GPT-4o Judge	3.25	4.13	4.64	5.52	17	9376
	+ RL w/ Qwen3-30B Judge	3.26	5.62	5.34	6.37	21	11529
	+ RL w/ GPT-4o Rubrics	3.43	4.19	5.08	6.12	18	10654
	+ RL w/ Qwen3-30B Rubrics	3.87	4.01	5.06	6.01	19	11142
Qwen3-4B-Instruct	Iterative Tool Loading	2.36	3.69	3.76	3.91	20	9045
	+ PTC	2.94	4.30	4.31	4.04	18	13667
	+ RL w/ Qwen3-30B Judge (Generic)	3.48	5.09	5.11	5.96	23	12815
	+ PTC + RL w/ Qwen3-30B Judge (Generic)	3.91	4.33	4.95	4.85	20	12951
	+ RL w/ Qwen3-30B Judge (Rubrics)	4.03	3.78	4.54	4.87	20	11151
	+ PTC + RL w/ Qwen3-30B Judge (Rubrics)	4.15	4.2	4.52	4.66	18	13400
Qwen3-30B-Instruct	Iterative Server Loading	2.51	3.83	4.31	3.17	28	9849
	+ RL w/ Qwen3-30B Judge (Rubrics)	4.08	3.68	4.45	3.68	21	9575
	Iterative Tool Loading	2.62	4.17	4.66	5.22	44	15334
	+ RL w/ Qwen3-30B Judge (Rubrics)	4.19	4.17	4.85	5.8	24	12936
	Iterative Tool Loading + PTC	3.11	3.51	5.03	4.58	18	16234
	+ RL w/ Qwen3-30B Rubrics	4.44	4.36	4.97	5.83	21	12201

Table 4: MCPBench Performance comparison across model families, training variants, and reward configurations. Scores correspond to evaluation metrics reported in the main text.

Model	Variant	Task Fulfillment	Grounding	Tool Appropriateness	Parameter Accuracy	Avg Turns	Avg Tokens
Kimi-K2 Thinking	All Tools Loaded	4.36	4.90	5.01	4.46	21	16025
	Iterative Server Loading	4.07	4.78	4.27	4.15	27	16554
	Iterative Tool Loading	3.70	4.17	3.96	3.01	30	12349
Qwen2.5-7B-Instruct	Iterative Server Loading	2.82	2.87	3.78	3.95	19	7778
	+ RL w/ GPT-4o Judge	2.94	3.05	4.22	4.18	17	7076
	+ RL w/ Qwen3-30B Judge	3.03	4.13	4.68	4.40	20	9896
	+ RL w/ GPT-4o Rubrics	3.52	3.58	4.31	4.39	15	7924
	+ RL w/ Qwen3-30B Rubrics	3.46	3.50	4.28	4.47	17	8756
Qwen3-4B-Instruct	Iterative Server Loading	3.16	3.87	4.51	4.23	18	6220
	+ RL w/ GPT-4o Judge	3.64	4.60	5.14	6.02	12	6951
	+ RL w/ Qwen3-30B Judge	3.09	4.32	4.77	4.73	21	9530
	+ RL w/ GPT-4o Rubrics	3.91	4.33	5.32	5.55	13	7175
	+ RL w/ Qwen3-30B Rubrics	4.12	4.09	5.14	5.13	15	8114
Qwen3-4B-Instruct	Iterative Tool Loading	2.96	3.57	3.85	3.17	17	7062
	+ PTC	2.76	3.50	3.83	2.70	14	7846
	+ RL w/ Qwen3-30B Judge	3.73	5.17	5.50	5.47	14	7391
	+ PTC + RL w/ Qwen3-30B Judge	3.79	4.26	4.72	3.61	18	11491
Qwen3-30B-Instruct	Iterative Tool Calling + PTC	3.06	3.29	4.50	3.17	21	14905
	+ PTC + RL w/ Qwen3-30B Rubrics	4.62	3.86	4.22	3.34	15	7488

Table 5: ATLAS-Test Performance comparison across model families, training variants, and reward configurations. Scores correspond to evaluation metrics reported in the main text.

# Servers	Difficulty	Train	MCPBench-Test	ATLAS-Test
1 Server	Easy	0	–	–
	Medium	36	–	49
	Hard	121	56	–
	Total	157	56	49
2 Server	Easy	44	–	5
	Medium	55	–	13
	Hard	29	30	12
	Total	128	30	30
3 Server	Easy	0	–	–
	Medium	9	–	21
	Hard	12	18	–
	Total	21	18	21
All	Total	304	104	100

Table 6: Task distribution across training and evaluation sets, stratified by number of MCP servers and task difficulty.

D DATASET

D.1 MCP TASKS

The task consists of answering a complex, multi-part, evidence-grounded user query by orchestrating multiple Model Context Protocol (MCP) servers, each exposing domain-specific tools and structured data sources. The system must decompose the user request, identify which sub-queries require external retrieval or computation, selectively invoke the appropriate MCP servers, and synthesize the results into a coherent response.

MCPBench. MCPBench consists of **28 MCP servers** exposing a total of **257 tools**. These servers span a broad range of real-world domains, including search, scientific computing, geospatial reasoning, finance, health, and knowledge retrieval. The full list of servers is: *OpenAPI Explorer, Unit Converter, Wikipedia, Google Maps, Bibliomantic, BioMCP, Call for Papers, Car Price Evaluator, Context7, DEX Paprika, FruityVice, Game Trends, Huge Icons, Hugging Face, Math MCP, NixOS, OSINT Intelligence, Reddit, National Parks, Medical Calculator, Metropolitan Museum, Movie Recommender, NASA Data, OKX Exchange, Paper Search, Scientific Computing, Weather Data, and Time MCP*.

ATLAS Server Extensions. To evaluate scalability beyond MCPBench, we augment the environment with **11 additional servers** exposing **57+ tools**. These servers introduce capabilities such as local computation, persistent memory, filesystem access, structured search, and stepwise reasoning. The added servers are: *calculator, cli-mcp-server, filesystem, knowledge-graph-memory, open-library, osm-mcp-server, pubmed, airbnb, duckduckgo-search, howtocook-mcp, and sequential-thinking*.

D.2 TRAINING SET AND BENCHMARKS

Motivation. The original MCPBench dataset provides **104 tasks**, which we found insufficient for stable reinforcement finetuning. Moreover, because MCPBench tasks are synthetically generated, a non-trivial fraction exhibit low task fulfillment even for strong agents, resulting in noisy learning signals. To address this, we construct a higher-quality training set while preserving MCPBench-style task structure and difficulty.

Training Set. Our final training set contains **304 tasks**. Tasks are synthetically generated using the MCPBench task generation pipeline with `o4-mini`, varying both *server combinations* and *task difficulty*. Difficulty is explicitly stratified into **Easy**, **Medium**, and **Hard**, corresponding to tasks with approximately **2–3**, **4–5**, and **6+** distinct requirements, respectively. This curriculum-style variation encourages stable learning across long-horizon, compositional tasks.

To ensure task quality, we apply the following automated filters using `o4-mini` as an LLM judge: (i) *Task solvability* (the extent to which the task can be completed using the available tools) must exceed **9**, and (ii) *Utility* (real-world usefulness) must exceed **8**.

Solvability Verification. To further ensure that retained tasks are genuinely solvable, all candidate tasks are executed using **Kimi-K2 Instruct** (1T parameters), a frontier model with strong tool understanding and long-horizon reasoning capabilities. Only tasks achieving a *task fulfillment score* greater than **5** under Kimi-K2 are retained for training.

Benchmarks. We evaluate models on two benchmarks:

- **In-distribution (ID):** the original **104 MCPBench tasks**, covering combinations of the 28 MCPBench servers.
- **Out-of-distribution (OOD): 100 tasks** constructed using **11 new additional servers** not observed during training, to assess generalization to novel tool compositions.

Together, these benchmarks enable controlled evaluation of both in-distribution performance and robustness to unseen server combinations.

E AUTOMATED RUBRIC GENERATION

E.1 EXAMPLES

User Query.

I'm gearing up for a new livestream and totally lost on what's big on Steam at the moment. Could you find out which game is currently dominating the charts? Once you've got that, hop over to its subreddit and see which discussions are really popping off say the top five hottest threads and tell me each thread's title and how many comments it's racking up. I need the real numbers and names, not just guesses, so I can plan my content around what people are actually talking about. Please ensure all findings are supported by concrete data and verifiable sources. I need specific numbers and evidence, not generalizations.

Task Requirements.

Using the provided Game Trends and Reddit tools, complete the following two requirements:

1. Retrieve the current top trending game on Steam.
2. Fetch the top 5 hot threads from the top trending game's dedicated subreddit (using the game name as the subreddit), and list each thread's title and number of comments.

Output Format:

```
{
  "game": "<top_trending_game_name>",
  "threads": [
    { "title": "<thread_title>",
      "comments": <comment_count> },
    ...
    // up to 5 entries
  ]
}
```

Auto-Generated Rubric.

1. **Criteria Name:** Task Fulfillment and Quality
Criteria Description: Accurate determination of the game currently dominating Steam's charts: the agent should select a sensible Steam chart to interpret "dominating the charts" (e.g., Top Sellers, Most Played/CCU, or Trending), justify the choice briefly, and correctly identify the current #1 title using its exact official name.
Weight: 10
2. **Criteria Name:** Task Fulfillment and Quality
Criteria Description: Data recency and currency: the agent must ensure the Steam and Reddit results reflect the current moment by making fresh tool calls and reporting retrieval timestamps (with timezone). The response should acknowledge that popularity and comment counts can change rapidly and reflect the state at the time of retrieval.
Weight: 10
3. **Criteria Name:** Task Fulfillment and Quality
Criteria Description: Correct subreddit identification for the chosen game: the agent should find the game's dedicated subreddit, handling naming variants (e.g., abbreviations, appended "Game" or "Official") and avoiding unrelated or generic subreddits. If multiple candidates exist, it should select the official/most active community and briefly justify the choice.
Weight: 9
4. **Criteria Name:** Task Fulfillment and Quality
Criteria Description: Complete and accurate retrieval of the top five "hot" threads: the agent should fetch exactly five posts from the subreddit's "hot" ranking, ensure they are unique, and handle stickied/moderation posts appropriately (include/exclude with an explicit rationale aligned with "hottest discussions").
Weight: 9

-
5. **Criteria Name:** Task Fulfillment and Quality
Criteria Description: Precision and verifiability of thread details: each thread must include the exact title and current number of comments as returned by the tool at retrieval time, along with direct permalinks to the posts and a link to the underlying Steam chart/source. No estimates or paraphrases; numbers and names must be exact and checkable.
Weight: 8
 6. **Criteria Name:** Tool Appropriateness
Criteria Description: Appropriate Steam data source: the agent should use a tool capable of providing up-to-date Steam chart data (e.g., official/top sellers/most played/trending endpoints or a trusted Steam charts tool). It should avoid outdated summaries or generic web search when a dedicated tool exists, and use a reasonable fallback only if the primary tool fails.
Weight: 7
 7. **Criteria Name:** Tool Appropriateness
Criteria Description: Appropriate Reddit retrieval method: the agent should use a Reddit API/tool that supports sorting by “hot” and limiting results to five. It should not scrape HTML when a Reddit tool is available and should respect rate limits/authentication requirements as applicable.
Weight: 8
 8. **Criteria Name:** Tool Appropriateness
Criteria Description: Use of auxiliary discovery tools for subreddit resolution: when the subreddit name does not exactly match the game title, the agent should leverage search/directory tools to find the official or most active subreddit and validate it via indicators like subscriber count, description, and recent activity.
Weight: 8
 9. **Criteria Name:** Tool Grounding
Criteria Description: All claims grounded in tool outputs: the identified game, subreddit, thread titles, and comment counts must directly match fields returned by the tools (e.g., “title”, “num_comments”). The agent should avoid hallucinations and ensure that any quoted numbers or names are traceable to the tool response.
Weight: 6
 10. **Criteria Name:** Tool Grounding
Criteria Description: Cross-step consistency and reconciliation: the agent should consistently use the game name returned by the Steam tool when searching Reddit and explicitly reconcile any mismatches (e.g., subreddit uses a shortened name). Any discrepancies should be explained with supporting evidence from the tools.
Weight: 7
 11. **Criteria Name:** Parameter Accuracy
Criteria Description: Reddit tool call parameters: the agent should pass the correct subreddit identifier, set `sort=hot`, `limit=5`, and include any necessary flags (e.g., handling NSFW if relevant). It must correctly interpret `num_comments` (not `score/upvotes`) as the comment count to report.
Weight: 5
 12. **Criteria Name:** Parameter Accuracy
Criteria Description: Steam tool call parameters: the agent should select the appropriate endpoint/parameters to represent “dominating the charts” (e.g., Top Sellers or CCU/Trending, `region=global/platform=PC` as appropriate), and document the chosen chart mapping. If multiple endpoints exist, parameters should unambiguously target the current chart leader.
Weight: 6

F PROMPTS

F.1 RUBRIC GENERATION PROMPT

In this section we show the prompt used by frontier LLM (GPT-5) to generate rubrics for the SLM Judge.

You are an expert in generating rubrics, given a user query, for evaluating the performance of AI agents on MCP based tool calling tasks. Your rubrics should be detailed and evidence-based and tailored to the specific task described in the `USER_QUERY`.

You will be given a `USER_QUERY` describing the task assigned to the agent and a `SOLUTION`, which demonstrates one of the possible approaches to solving the task. Based on this, generate a comprehensive rubric for evaluating the agent's trajectory on the task. There could be multiple ways to solve the task, so ensure your rubric is general enough to accommodate different valid approaches while still being specific to the task requirements. Use the `SOLUTION` to understand the task requirements and the challenges involved and do not base your rubric solely on the `SOLUTION` provided. You can always use insights from the `SOLUTION` to inform your rubric generation.

You are also given the `CRITERIA_CATEGORIES` and `SCORING INSTRUCTIONS` to guide your rubric generation. `CRITERIA_CATEGORIES` define the key dimensions to evaluate, and `SCORING INSTRUCTIONS` provide guidance on how to assign weights to each criterion.

`CRITERIA_CATEGORIES`:

1. Task Fulfillment and Quality : Evaluate how well the agent met the task requirements and whether all the subtasks were completed satisfactorily.
2. Tool Appropriateness : Evaluate whether the tools selected by the agent were suitable for the tasks they were intended to solve.
3. Tool Grounding : Evaluate how well the agent's reasoning is grounded in the tool outputs and whether the agent effectively utilized the information provided by the tool response
4. Parameter Accuracy : Evaluate the accuracy and completeness of the parameters used in the tool calls.

`SCORING INSTRUCTIONS`:

- 1-3: If the criterion is of low importance for evaluating the `USER_QUERY` and it only evaluates minor aspects of the task, that need not be necessary, but would be a good to have properties while executing the task.
- 4-6: If the criterion is of moderate importance for evaluating the `USER_QUERY` and it evaluates important aspects of the task but not critical for overall success. These could be aspects that improve the task execution but are not essential.
- 7-8: If the criterion is of high importance for evaluating the `USER_QUERY` and it significantly impacts the success of the task and its outcomes. These could be aspects that greatly enhance the quality or effectiveness of the task execution.
- 9-10: If the criterion is of critical importance for evaluating the `USER_QUERY` and it is essential for the successful completion of the task without which the task would fail.

As you response return a list of JSON objects with the following structure:

```
{
  "criteria_name": "name any one of the four CRITERIA_CATEGORIES",
  "criteria_description": "detailed description of what this criteria evaluates with respect to the USER_QUERY. Keep this focused on the specific task described in the USER_QUERY.",
  "weight": "return a number from 1-10, both included indicating the importance of this criteria for evaluating the USER_QUERY and use the SCORING INSTRUCTIONS to guide your weight assignment."
}
```

INSTRUCTIONS:

- It is important that Task Fulfillment and Quality is always included and given the highest weight among the rubrics
- Tool Appropriateness is the second most important criteria and should be generally given the second highest weight among the rubrics
- Always return a list of JSON objects as described above with 12 rubrics out of which follow the below distribution in the same order:
- 5 rubrics for Task Fulfillment and Quality
- 3 rubrics for Tool Appropriateness
- 2 rubrics for Tool Grounding
- 2 rubrics for Parameter Accuracy
- Ensure that each of the 12 rubrics corresponds to a unique aspect of the USER_QUERY and does not overlap with other rubrics.
- Ensure the descriptions are specific to the USER_QUERY and provide clear guidance on what to evaluate for each criterion.
- Return the list between <RUBRICS >[{} , {} , ...]</RUBRICS >tags only.

USER_QUERY : {user_query}

SOLUTION : {solution}

F.2 RUBRIC EVAL PROMPT

The following prompt was used by the SLM Judge to generate rewards for a given trajectory.

You are an expert evaluator for assessing the performance of AI agents on MCP based tool calling on a set of pre-defined rubrics. Given a USER_QUERY describing the task assigned to the agent, a RUBRICS list defining the evaluation criteria, and a TRAJECTORY representing the agent's actions and outputs during task execution, your goal is to return a number between 0 and 1 for each rubric based on the degree to which the agent met the expectations outlined in that rubric based on the below criteria :

- 0.1–0.3: 10-30% of trajectory steps were correctly executed according to the rubric.
- 0.4–0.6: 40-60% of trajectory steps were correctly executed according to the rubric.
- 0.7–0.8: 70-80% of trajectory steps were correctly executed according to the rubric.
- 0.9–1.0: 90-100% of trajectory steps were correctly executed according to the rubric.

USER_QUERY : user_query

RUBRICS : rubrics

TRAJECTORY : trajectory

In your response, return a list of numbers between 0 and 1 corresponding to each rubric in the RUBRICS list in the same order as listed.

You must return a list of 12 decimal numbers between <EVAL_SCORES >and </EVAL_SCORES >tags only as given below:

<EVAL_SCORES >[0.1, 0.8, 1, ..., 0]</EVAL_SCORES >

F.3 BASE JUDGE PROMPT

The following is the prompt employed by the o4-mini LLM Judge to generate all the evaluation metrics.

You are a STRICT evaluator. Your role is to critically assess performance with HIGH STANDARDS.

IMPORTANT: The average score across all evaluations should be around 4-5, NOT 7-8.

You must assign scores ****only based on evidence**** from the task, solution, and tool usage. Your evaluation should be:

- Extremely Critical (assume mediocre performance unless proven otherwise)
- Evidence-based (require strong proof for scores above 5)
- Conservative (when in doubt, score lower - aim for 4-5 average)

CRITICAL FORMAT RULES:

- DO NOT penalize for output format (JSON, text, etc.) unless the TASK PRESENTED TO AGENT explicitly requires it
- If the task presented to agent says "provide information" without specifying format, ANY readable format is acceptable
- Only deduct points for format if the task explicitly states "return as JSON" or "format as table" etc.
- Focus on CONTENT correctness, not presentation style

—
****TASK PRESENTED TO AGENT****: "{task}"

****CONCRETE TASK REFERENCE (For evaluation context only)****:
{concrete_task_description}

Note: The agent did NOT see this concrete version. It only saw the task above.
The task visible for the agent is the fuzzy version of the concrete task.
This reference helps assess actual task completion but is not the sole criterion.
The agent's interpretation of the fuzzy task may differ but still be valid.

FORMAT REMINDER: If the concrete task mentions JSON but the TASK PRESENTED TO AGENT doesn't explicitly require it,
DO NOT penalize for not using JSON format. Only the task presented to agent's requirements matter for format.

****EXECUTION SUMMARY****:

{execution_summary}

****FINAL SOLUTION****: "{final_solution}"

****TOTAL ROUNDS****: {total_rounds}

—
Task Completion Rubric (1–10 per subdimension)

1. ****Task Fulfillment and Quality****

- 1–3: Perfectly completes 10-30% of requirements.
- 4–6: Perfectly completes 40-60% of requirements.
- 7–8: Perfectly completes 70-80% of requirements.
- 9–10: Perfectly completes 90-100% of requirements.

NOTE: Requirements come from the task present to agent only. Format (JSON/text) is NOT a requirement unless explicitly stated in the task present to agent.

3. ****Grounding****

- 1–3: 10-30% of claims are perfectly grounded in tool outputs.
- 4–6: 40-60% of claims are perfectly grounded in tool outputs.
- 7–8: 70-80% of claims are perfectly grounded in tool outputs.

- 9-10: 90-100% of claims are perfectly grounded in tool outputs.

Tool Usage Rubric (1-10 per subdimension)

1. **Tool Appropriateness**

- 1-3: 10-30% of tools were perfectly selected for their subtasks.
- 4-6: 40-60% of tools were perfectly selected for their subtasks.
- 7-8: 70-80% of tools were perfectly selected for their subtasks.
- 9-10: 90-100% of tools were perfectly selected for their subtasks.

3. **Parameter Accuracy**

- 1-3: 10-30% of tool calls have perfectly accurate and complete parameters.
- 4-6: 40-60% of tool calls have perfectly accurate and complete parameters.
- 7-8: 70-80% of tool calls have perfectly accurate and complete parameters.
- 9-10: 90-100% of tool calls have perfectly accurate and complete parameters.

PERCENTAGE-BASED SCORING SYSTEM:

How to Calculate Scores:

For each dimension, calculate the DEFECT RATE: - Defect Rate = (Number of Issues / Total Opportunities) × 100%

Then map defect rate to score:

- 0-10% defects → Score 9-10 (Excellent to Perfect)
- 10-30% defects → Score 7-9 (Good performance)
- 30-50% defects → Score 5-7 (Average performance)
- 50-70% defects → Score 3-5 (Poor performance)
- 70-100% defects → Score 0-3 (Failed)

How to Score:

1. When evaluating percentages, be EXTREMELY STRICT about what counts as "perfectly executed"
2. "Perfectly" means ALL of the following must be true:
 - Correct tool selection (not just "works" but OPTIMAL choice)
 - Complete and accurate parameters (not just valid, but IDEAL)
 - Zero redundancy (no repeated or unnecessary calls)
 - Proper error handling (graceful recovery from ANY failure)
 - Efficient execution (parallel when possible, minimal rounds)
 - Concise output (no verbose explanations unless requested)
3. If ANY of the above is missing, that portion is NOT perfectly executed (counts as 0%)
4. Example: Task completed correctly but with 1 redundant call = that portion is 0% perfect

KEY PRINCIPLES:

1. ALWAYS calculate as percentage, NOT absolute numbers
2. 10 errors in 100 calls (10%) = same score as 1 error in 10 calls (10%)
3. Consider the OPPORTUNITY COUNT for each dimension: - Tool calls: How many total calls were made? - Parameters: How many total parameters across all calls? - Claims: How many factual statements were made?

CRITICAL: Apply the STRICTEST interpretation of "perfectly executed". If there's ANY doubt, score lower.

CONCRETE SCORING EXAMPLES WITH PROPORTIONS:

Task Fulfillment:

- Completed 19/20 requirements (5% defect rate) = Score 9
- Completed 16/20 requirements (20% defect rate) = Score 8
- Completed 12/20 requirements (40% defect rate) = Score 6
- Completed 8/20 requirements (60% defect rate) = Score 4

Tool Appropriateness:

- 19/20 tools optimal (5% defect rate) = Score 9
- 16/20 tools optimal (20% defect rate) = Score 8
- 12/20 tools optimal (40% defect rate) = Score 6
- 8/20 tools optimal (60% defect rate) = Score 4

Grounding:

- 19/20 claims supported by evidence (5% unsupported) = Score 9
- 16/20 claims supported by evidence (20% unsupported) = Score 8
- 12/20 claims supported by evidence (40% unsupported) = Score 6
- 8/20 claims supported by evidence (60% unsupported) = Score 4

Parameter Accuracy: - 95/100 parameters perfect (5% defect rate) = Score 9

- 80/100 parameters perfect (20% defect rate) = Score 8
- 60/100 parameters perfect (40% defect rate) = Score 6
- 40/100 parameters perfect (60% defect rate) = Score 4

FORMAT NOTE: Text output when JSON not required in the task present to the agent = NO PENALTY (0% defect) FORMAT NOTE: Missing JSON when explicitly required in the task present to the agent = Count as failed requirement

Remember: Most real-world executions should score 4-6. Scores of 8+ should be EXCEPTIONAL.

FINAL REMINDER BEFORE SCORING:

- Default to 4-5 unless you have strong evidence for higher
- Count ONLY truly perfect executions toward the percentage
- Be your most critical self - find flaws first, then acknowledge successes
- If you're considering a score above 7, re-examine for ANY imperfection
- Server count is IRRELEVANT - using more servers is NOT better

Please score based on COMPLETION PERCENTAGES and PROPORTIONAL SUCCESS, not absolute numbers. Return your evaluation scoring and reasoning in this exact JSON format:

```
{
  "task_fulfillment_reasoning": "Explain how well the agent fulfilled the detailed task objectives, referencing specific content from the CONCRETE TASK DESCRIPTION and what percentage was completed.",
  "grounding_reasoning": "Explain how well the agent's outputs were grounded in actual tool results versus unsupported claims.",
  "tool_appropriateness_reasoning": "Explain whether the tools selected were appropriate for each subtask requirement.",
  "parameter_accuracy_reasoning": "Explain the accuracy and completeness of parameters used in tool calls, noting any missing required parameters or incorrect values.",
```

```
"task_fulfillment": X,
"grounding": X,
```

```
"tool_appropriateness": X,
"parameter_accuracy": X,
```

```
"dependency_awareness": X,  
"parallelism_and_efficiency": X,  
}
```

Return **only** the JSON object.

F.4 AGENT PROMPTS

F.4.1 ISL PROMPT

You are a strategic decision-making expert that makes tool calls to solve a task. You will use the provided MCP servers and their tools to plan and execute the given task effectively.

You will be given a task and a list of MCP servers, each with specific capabilities. Follow these steps carefully:

Perform Decision and Planning:

1. Assess if the task is fully completed. If it is, return a final answer enclosed in `<answer> ... </answer>` tags.
2. If not, decide if executing tools is valuable.
3. If executing tools is not valuable, return the best possible final answer enclosed in `<answer> ... </answer>` tags.
4. If further tool calls are needed, follow the Execution Planning steps below and return tool call enclosed in `<tool_call>... </tool_call>` tags.

Execution Planning:

1. Create a step-wise plan to accomplish the task using the available MCP servers and their tools.
2. Infer the possible description or capability of each MCP server from its name.
3. Identify which MCP server(s) are most relevant based on these inferred roles.
4. Select the best server(s) to accomplish the task.
5. If tools of the server are unknown, use the tool 'fetch_tools' to get the list of tools available within that server.
6. Choose the most appropriate tool(s) from the selected server(s) to address the task.
7. Update your plan based on the results obtained from tool executions.

When planning next steps:

- Build on previous results.
- Reason the tool to use.

Response format should be one of the following:

- If the task is completed or no further tool calls are valuable:

```
<think>...your reasoning (check task completion)...</think>  
<answer>...final answer...</answer>
```

- If further tool calls are needed:

```
<think>...your reasoning...</think>  
<tool_call>...tool call...</tool_call>
```

Note:

- For MCP server, tool name is "server:tool_name".
- Always start with `<think>...</think>` and each response should contain only one `<think>...</think>`.
- Include the decision and planning steps in your reasoning `<think>...</think>`.
- All tool calls must be wrapped in `<tool_call>...</tool_call>`.
- **AVOID REDUNDANT CALLS:** Don't repeat successful tools unless specifically needed.
- All final answers must be wrapped in `<answer>...</answer>`.

- Before giving the final answer, ensure all aspects of the task are addressed.

F.4.2 ITL PROMPT

You are a strategic decision-making expert for a multi-tool AI agent. You will use the provided MCP servers and their tools to plan and execute the given task effectively.

You will be given a task and a list of MCP servers, each with specific capabilities. Follow these steps carefully:

Perform Decision and Planning:

1. Assess if the task is fully completed. If it is, return a final answer enclosed in `<answer> ... </answer>` tags.
2. If not, decide if executing tools is valuable.
3. If executing tools is not valuable return the best possible final answer enclosed in `<answer> ... </answer>` tags.
4. If further tool calls are needed, follow the Execution Planning steps below and return tool call enclosed in `<tool_call> ... </tool_call>` tags.

Execution Planning:

1. Create a step-wise plan to accomplish the task using the available MCP servers and their tools.
2. Infer the possible description or capability of each MCP server from its name.
3. Identify which MCP server(s) are most relevant based on these inferred roles.
4. Select the best server(s) to accomplish the task.
5. If tools of the server are unknown, use the tool 'fetch_tools' to get the list of tools available within that server.
6. Choose the most appropriate tool(s) from the selected server(s) to address the task.
7. Update your plan based on the results obtained from tool executions.

When planning next steps:

- Build on previous results.
- Reason the tool to use.

Note:

- For MCP server tool name is "server:tool_name".
- All tool calls must be wrapped in `<tool_call>...</tool_call>`.
- AVOID REDUNDANT CALLS: Don't repeat successful tools unless specifically needed
- All final answers must be wrapped in `<answer>...</answer>`.
- Before giving final answer, ensure all aspects of the task are addressed.

Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools> { "type": "function", "function": { "name": "get_selected_tools", "description": "A tool for fetching schema of selected tools for a particular MCP server.", "parameters": { "type": "object", "properties": { "server": { "type": "string", "description": "The MCP server to fetch tools from." }, "tools": { "type": "array", "description": "List of tools to fetch" } }, "required": [ "server", "tools" ] } } } </tools>
```

For each function call, return a json object with function name and arguments within `<tool_call></tool_call>` XML tags: `<tool_call> { "name": <function-name>, "arguments": <args-json-object> } </tool_call>`

F.4.3 PTC PROMPT

You are a reasoning language model that can reach precise answers through careful reasoning and tool use when needed.

Tools:

You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools>
{
  "type": "function",
  "function": {
    "name": "code_interpreter",
    "description": "A tool for executing Python code in Jupyter notebook style cells, state is maintained across the calls",
    "parameters": {
      "type": "object",
      "properties": {
        "code": {
          "type": "string",
          "description": "The raw Python code string to be executed."
        }
      }
    },
    "required": ["code"]
  }
}
</tools>
```

Instructions:

1. Always enclose your internal reasoning process within `<think>...</think>` blocks.
2. To invoke the Python `code_interpreter` tool, return a JSON object containing the tool name and the code argument within `<tool_call>...</tool_call>` tags.
3. Once you have sufficient information, provide your final answer to the user.
4. Use the code interpreter tool to execute Python code and invoke `MCP`Server methods as needed.

MCP Python Coding Solving Rules:

1. Come up with a brief plan of the steps required to complete the user-specified MCP task with the servers and tools given.
2. While writing Python tool calls, first import the `MCP`Server object from the `MCP`Bench package and initialize an `MCP`Server object with the desired server name.
3. The first Python call should always fetch the expected argument formats and output formats for the tools you will need. Use:
`mcpobject.get_tools_info(['tool1', 'tool2', ...])`
4. Only after understanding the tool schemas, write the code to implement your plan.
5. If you encounter errors due to format issues, call `mcpobject.get_tools_info(['tool1', 'tool2', ...])` again.

Example tool calls:

Fetching tool details to get schema for the relevant tools:

```
<tool_call> { "name": "code_interpreter", "arguments":
{ "code": "from MCPBench import MCPServer; time_mcp =
MCPServer('Time MCP'); time_mcp.get_tools_info(['get_current_time', 'convert_time'])"
} } </tool_call>
```

Calling a tool after fetching the schema:

```
<tool_call> { "name": "code_interpreter", "arguments": {  
  "code": "time_mcp.get_current_time(timezone='America/New_York')"  
} } </tool_call>
```