

# ToolGym: an Open-world Tool-using Environment for Scalable Agent Testing and Data Curation

Ziqiao Xi\* Shuang Liang\* Qi Liu Jiaqing Zhang Letian Peng  
Fang Nan Meshal Nayim Tianhui Zhang Rishika Mundada  
Lianhui Qin Biwei Huang Kun Zhou  
UC San Diego

## Abstract

Tool-using LLM agents still struggle in open-world settings with large tool pools, long-horizon objectives, wild constraints, and unreliable tool states. For scalable and realistic training and testing, we introduce an open-world tool-using environment, built on 5,571 format unified tools across 204 commonly used apps. It includes a task creation engine that synthesizes long-horizon, multi-tool workflows with wild constraints, and a state controller that injects interruptions and failures to stress-test robustness. On top of this environment, we develop a tool select–then–execute agent framework with a planner–actor decomposition to separate deliberate reasoning and self-correction from step-wise execution. Comprehensive evaluation of state-of-the-art LLMs reveals the misalignment between tool planning and execution abilities, the constraint following weakness of existing LLMs, and DeepSeek-v3.2’s strongest robustness. Finally, we collect 1,170 trajectories from our environment to fine-tune LLMs, achieving superior performance to baselines using 119k samples, indicating the environment’s value as both a realistic benchmark and a data engine for tool-using agents. Our code and data will be publicly released.

## 1 Introduction

Owing to scaling laws, large language models (LLMs) have exhibited strong instruction following, reasoning, and planning abilities (Kaplan et al., 2020; Wei et al., 2022; Bubeck et al., 2023), catalyzing the rapid rise of autonomous LLM-based agents (Xi et al., 2025; Park et al., 2023). To extend these agents beyond text generation, tool use has become a key capability (Mialon et al., 2023a; Schick et al., 2023). By connecting an LLM to external tools, an agent can take actions in the world—querying services, operating software, and executing real workflows. With carefully designed

toolsets and interfaces, tool-using agents have already shown strong practical competence, from information seeking to complex coding tasks, and are increasingly positioned to offload repetitive or heavy workloads from humans.

Despite rapid progress, existing tool-using agents remain far from human performance in real-world applications, where the decision space is large and noisy, task requirements are often under-specified or fuzzy, and success depends on satisfying many constraints (Zhou et al., 2024; Mialon et al., 2023b; Liu et al., 2024b). In practice, an agent must choose the right tool and invoke it correctly, and adaptively switch among a large tool pool as the situation evolves. Meanwhile, it also needs to handle complex user preferences (*e.g.*, using multiple services, making explicit trade-offs) and unreliable tool states (*e.g.*, flaky servers, timeouts, rate limits, or expired credentials). Humans manage these challenges largely because we accumulate rich experience and commonsense from everyday interactions, learning how tools behave through repeated trial, error, and adjustment.

By contrast, today’s agents lack a realistic and scalable environment to test and acquire such experience, making it difficult to form a test-then-learn loop akin to how humans learn a new tool. As shown in Table 1, existing benchmarks and datasets are bottlenecked by limited tool scale or overly simplified settings that sidestep real-world constraints (to avoid unstable states and complicated configurations). As a result, they cannot fully capture the conditions that matter in practice, *e.g.*, long-horizon workflows over large tool libraries, open-ended tool selection under dense compositional requirements, and interactions with stateful services that may fail. Therefore, strong benchmark scores often correlate poorly with real user experience, and struggle to reflect the robustness and generalization of agents in real-world applications.

To address this gap, we build an open-world tool-

\* Equal contribution

Benchmark	# Apps / # Servers	# Tools	# Rounds	Fully Automatic	Wild Constraints	State Controller	Fuzzy Instructions
BFCL v4 (Patil et al.)	1 / None	2000+	3.8	✗	✗	✗	✗
ToolBench (Qin et al., 2024)	N/A	3451	3.7	✓	✗	✗	✗
AgentBench (Liu et al., 2024b)	N/A	18	11.9	✗	✗	✗	✗
ToolEyes (Ye et al., 2025)	N/A	568	4.5	✗	✗	✗	✗
StableToolBench (Guo et al., 2024)	N/A	3451	3.7	✓	✗	✓	✗
MCPEval (Liu et al., 2025)	8 / 14	121	N/A	✓	✗	✗	✗
LiveMCPBench (Mo et al., 2025)	36 / 70	527	5.6	✗	✗	✗	✗
MCP-Universe (Luo et al., 2025)	8 / 11	133	7.5	✗	✗	✗	✗
MCP-Bench (Wang et al., 2025)	20 / 28	250	44.2	✓	✓	✗	✓
Toolathlon (Li et al., 2025)	17 / 32	604	26.8	✗	✓	✓	✓
ToolAce (Liu et al., 2024a)	N/A	26507	1.6	✓	✗	✗	✗
Toucan (Xu et al., 2025)	179 / 495	2000+	2	✓	✗	✗	✗
<b>ToolGym (Ours)</b>	204 / 276	5,571	28.5	✓	✓	✓	✓

Table 1: Comparison of related work along realism-oriented axes: # Apps, # Tools, and # Rounds denote the average number of covered real-world apps, available tools and tool use rounds per task for each model. We additionally mark whether it supports automatic data synthesis for scaling up, testing wild constraints following, and flexible state controlling.

using environment, named ToolGym that enables scalable training and evaluation of an LLM’s ability to operate as a tool-using agent. As its foundation, we manually curate and validate 5,571 tools spanning 204 commonly used desktop applications and online services, and standardize them into a unified MCP format to support consistent invocation and benchmarking. On top of this tool base, we develop a task creation engine that generates long-horizon, multi-tool workflows with wild realistic constraints, allowing us to synthesize large-scale tasks that better resemble the complex requests users encounter in everyday life. Finally, we introduce a state-control mechanism that can flexibly inject interruptions and simulate real-world issues (*e.g.*, unstable responses or transient failures) during execution, making it possible to systematically test and ultimately improve the robustness of tool-using agents under non-ideal conditions.

Building on this environment, we design an agent framework that follows a simple tool select–then–execute loop. Concretely, the agent retrieves relevant tools from a large MCP library, invokes them with appropriate arguments, and iteratively updates its state as observations return. Because real-world tasks are typically long-horizon and error-prone, we further introduce a planner–actor decomposition. In this way, a planner performs deliberate reasoning, reflection, and self-correction to maintain global progress and resolve conflicts, while an actor focuses on efficient step-by-step execution in the open-world setting. This functional disentanglement improves the

agent’s ability to complete complex workflows, and provides a clearer lens to analyze how LLMs reason, adapt, and fail across long-horizon trajectories. Using our environment and framework, we test state-of-the-art LLMs and uncover few findings:

- All LLMs exhibit strong planning ability, but their execution ability is not aligned and cause significant gap in task successful rate;
- Constraint following, rather than tool invocation, is the dominant failure mode of LLMs;
- Deepseek-v3.2 achieves the best robustness, with a stronger recovery capability from intermediate unpredictable issues;
- Higher tool-using rate does not imply higher successful rate, as weaker models make more tool calls yet fail more due to poor reasoning;

Moreover, because our data generation pipeline is fully automated, the environment can also serve as a scalable engine for collecting high-quality trajectories to improve tool-using agents. We use it to collect 1,170 samples to fine-tune Qwen2.5-7B and Qwen3-8B, and experiments show that the resulting models achieve superior performance to baselines trained with 119k samples, highlighting the strong data quality enabled by our environment. Finally, thanks to the flexible evaluation protocol and open-world setting, we can systematically probe the decision-making and planning styles of different LLMs under realistic tool-using conditions.

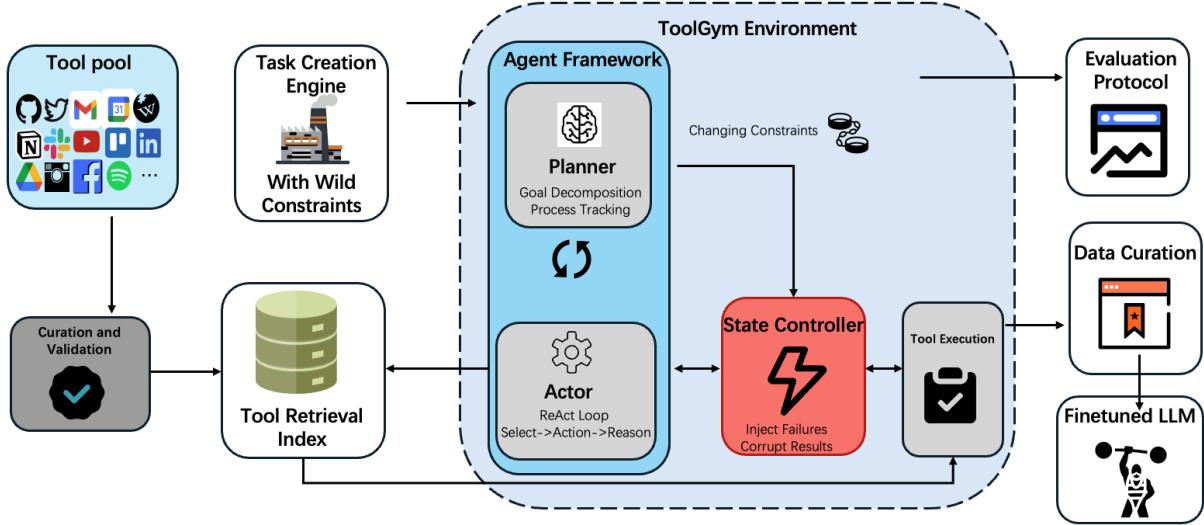


Figure 1: The overall framework of TOOLGYM. The pipeline begins with curating real-world MCP tools and synthesizing tasks with wild constraints (Left). The agent employs a Planner–Actor architecture to decompose long-horizon goals, where the Actor interacts with the environment via a State Controller. Crucially, the Controller intercepts tool calls to inject realistic failures (*e.g.*, timeouts, state changes) before execution on real servers. The resulting trajectories support both rigorous Evaluation and high-quality Data Curation for model training.

## 2 ToolGym

ToolGym is an open-world tool-use environment designed to test and improve LLM-based agents. It is built on a large-scale, extensible toolset that we manually curate and standardize under the MCP format. To ensure scalability, ToolGym provides a fully automated and interactive pipeline, including a task creation engine for synthesizing realistic workflows, an evaluation protocol for solution verification and judging, and a state controller for managing and intervening in intermediate agent states during execution.

### 2.1 Extensible Toolset Construction

To build the large-scale toolset in an efficient and extensible way, we unify the tool format using Model Context Protocol (MCP), select commonly used tools in daily human workflows, and validate them under controlled credentials. To support flexible tool use and realistic open-world scenarios, we additionally provide an MCP-based tool retrieval index that enables agents to search and manipulate tools from a massive library.

**MCP Server Selection.** We construct an executable tool universe by curating 5,571 tools from 276 MCP servers on Smithery<sup>1</sup>, covering 204 production-grade applications (*e.g.*, Gmail,

GitHub). Our selection follows a simple principle: we prioritize tools that are *commonly used*, *general-purpose*, and *popular* in everyday human work, so that the resulting toolset reflects realistic workflows (*e.g.*, documents, coding, search, and collaboration). We build this collection through a hybrid pipeline that combines automated registry crawling with manual augmentation, allowing us to scale coverage while preserving tool diversity and practical relevance. To keep the setting faithful to real-world APIs, we retain each tool’s native invocation interface and directly use its original argument schema. For servers requiring authentication, we provision dedicated virtual accounts without personal information and configure them with controlled credentials, enabling authorized execution while ensuring reproducibility across runs.

**Tool Validation.** The availability of the registry does not guarantee the executability of the tools. After filtering servers that fail basic MCP handshakes, we validate each individual tool under the authenticated evaluation personas using a three-stage pipeline: (i) *authenticated availability* ensuring the server can list and access the tool under the persona; (ii) *successful invocation* verifying the tool runs with schema-respecting arguments; and (iii) *usable responses* filtering tools that execute but return error payloads or non-actionable outputs. This process yields a reliable *working* toolset that

<sup>1</sup><https://smithery.ai>

can be safely used for downstream task synthesis and trajectory-based evaluation.

**Tool Retrieval Index.** Given the scale of our open-world tool registry, agents need an efficient mechanism for tool discovery, analogous to a search engine over APIs. We therefore index all validated tools as vector documents by concatenating their server identity, tool name, natural-language description, and schema signature. We embed these documents using BGE-M3 (Chen et al., 2024) and store them in a FAISS index (Johnson et al., 2021). At runtime, we provide a lightweight MCP service `search_tools(query, k)` on a locally deployed server, enabling agents to issue natural language queries and retrieve the top- $k$  relevant tools for on-demand loading and invocation.

## 2.2 Task Creation Engine

Our goal is to generate tasks that resemble what humans encounter in daily work: they are *long-horizon*, require *coordinated multi-tool use*, and include *wild constraints* (e.g., multi-source verification, and explicit trade-offs). To this end, we first sample a coherent and diverse candidate set of tools that can plausibly serve one workflow, and then synthesize the task that naturally uses these tools while considering diverse constraints.

**Tool Candidate Sampling.** We build a tool candidate set that is both semantically coherent (so the task is natural) and diverse across services (so it requires real orchestration). Concretely, we first randomly sample 1-3 *seed tools*, then form a retrieval query from their names and descriptions, and use our tool retrieval index to recall a larger set of semantically related tools. Next, we group recalled tools by their underlying servers/apps and uniformly sample across these groups (round-robin) to assemble the final candidate set. This step prevents the candidate set from being dominated by a single service and encourages cross-app workflows. We also control difficulty by sampling the final bundle size from a configurable range and enforcing a minimum number of distinct servers.

**Synthesizing Task with Constraints.** Given the selected tool candidate set, we first synthesize an initial task query by prompting an LLM to produce the task instruction with a unified goal that plausibly requires composing the candidate tools. Starting from this draft, we then run an iterative *check-then-revise* loop to progressively improve

realism and increase constraint density. In each round, we automatically evaluate the draft from two perspectives: (i) tool coverage: whether the request legitimately motivates using all or most candidate tools (supported by per-tool rationales); (ii) constraint quality: whether the constraints are diverse, non-trivial, and interact in ways that create long-horizon dependencies. If either criterion is not met, we generate targeted feedback and revise the prompt to re-synthesize a stronger query in the next iteration. The final task instruction includes the user query, a set of constraints, and grounded rationales describing each tool’s role, enabling long-horizon supervision and reliable evaluation.

## 2.3 Task Evaluation Protocol.

Given the complexity of our synthesized open-world tasks, we evaluate agent performance using an LLM-as-judge protocol with three state-of-the-art models: GPT-4o, GPT-5.1, and DeepSeek-V3.2. To stabilize evaluation, we aggregate their judgments via majority vote to produce the final label. This design reduces reliance on any single judge and improves robustness across different model families, perspectives, and cost regimes. For each episode, every judge receives a structured evaluation package containing the task query, the extracted constraint list, the agent’s final answer, and the full execution trajectory. Judges are instructed to assess two core axes: *task success* and *constraint satisfaction*. Beyond these default dimensions, our protocol is extensible: new prompts or metrics can be added to target specific properties such as robustness under injected failures or long-horizon planning quality. Overall, this protocol provides a scalable and reproducible way to assess multi-dimensional agent capabilities without relying on expensive human annotation.

## 2.4 State Controller

Real-world tool ecosystems are inherently unreliable, yet many benchmarks evaluate agents primarily on the idealized “happy path.” To systematically probe robustness under non-ideal conditions, we introduce a state controller: a middleware that can intervene during execution to create controlled failures and perturbations. Unlike random noise injection, the state controller applies *targeted* interventions that are reproducible and stress the agent’s recovery and adaptation abilities. Concretely, the state controller can inject controlled disruptions according to pre-defined policies, enabling fine-

grained robustness testing while keeping the task viable. We categorize interventions into three complementary control types:

- **Tool-level control:** Intercepts specific tool calls and injects realistic failures, *e.g.*, time-outs, rate limits, and temporary unavailability.
- **State-level control:** Modifies the intermediate execution state exposed to the agent, *e.g.*, delayed or corrupted tool results, changed service state (*e.g.*, session timeout), emulating common real-world drift and instability.
- **Constraint-level control:** Introduces or alters task constraints during execution (*e.g.*, a new preference or a changed deadline), testing whether the agent can maintain and re-optimize under evolving requirements.

This design creates a consistent adversity signal across models, and makes robustness measurable and comparable. A competent agent requires genuine resilience: it must detect the failure, revise its plan, select suitable alternatives, and complete the workflow while still satisfying constraints.

### 3 Agent Framework

The primary challenge in real-world scenarios is long-horizon decision making under high complexity, where agents must coordinate many interdependent sub-goals across dozens of tool-use steps, requiring sustained reasoning and planning. In this regime, purely zero-shot execution without external feedback or verification often fails due to premature termination: agents struggle to maintain coherent plans over long trajectories and to recover from intermediate errors. To address this, we introduce a *plan-then-action* framework that separates the agent into two roles, stabilizing execution through deliberate planning and iterative action.

#### 3.1 Agent Roles

We use the LLM to play two roles in the agent for long- and short-term decision making:

**The Actor.** The Actor functions as the execution engine following the ReAct paradigm (Yao et al., 2022). It operates in a step-wise manner that invokes the tool and then executes it based on the context and current state.

**The Planner.** The planner maintains a holistic view of the task and continuously tracks the Actor’s execution progress. Before execution begins, the planner explicitly decomposes the user intent into a reference sub-goal graph. During execution, it compares the Actor’s trajectory against this graph. If the Actor deviates, the planner intervenes with feedback to force alignment.

#### 3.2 Execution Workflow

**Task Decomposition and Planning.** The planner receives the task instruction and explicitly constructs a reference sub-goal graph. By maintaining the explicit plan, we enable active control and execution guidance when necessary.

**Tool Selection and Execution.** Driven by an implicit sub-goal, the Actor executes the following loop within the React Paradigm: (1) Reasoning and Tool Selection: It articulates a rationale to translate the sub-goal into a concrete search strategy, then queries the tool index to select the most relevant APIs; (2) Tool Invocation: It performs tool calling based on the schemas of the retrieved tools, and generates the final API call code for execution.

**Planner Feedback.** To enforce long-horizon consistency, after tool execution, the planner will verify the semantic validity and sufficiency between the execution result and the current sub-goal. Then, it marks the corresponding sub-goal status nodes as *Complete* or *Pending*. Once the last node is complete, the planner will terminate the control and return the final result. Besides, we also add the maximum turn limit and repeated laziness limit (refuse to use tools) as the termination conditions to avoid meaningless running of the agent.

### 4 Ability Evaluation with ToolGym

#### 4.1 Experimental Setup

We evaluate LLMs within TOOLGYM to simulate the demands of the open-world tool ecosystem. Moving beyond isolated function calling, our protocol targets *long-horizon orchestration*: agents must autonomously discover tools from the massive registry, adhere to *wild constraints*, and recover from unexpected failures to resolve complex objectives without pre-defined trajectories. To probe the upper bounds of agentic capability, we utilize our TASK CREATION ENGINE to synthesize a specialized evaluation set of 50 scenarios. Each task is evaluated with *pass@3* to account for stochasticity in

agent behavior and tool execution. Unlike standard benchmarks constrained by guaranteed solvability, this set is designed as a *stress testing*. We explicitly relax the constraint of solvability to prioritize combinatorial difficulties, introducing conflicting constraints and inducing long trajectories to identify the breaking point of current frontier models.

## 4.2 Results Analysis.

Table 2 presents the comparative performance across different model tiers. The details of evaluation metrics are in Appendix. Based on the metrics, we observe systematic performance differences across models, accompanied by distinct behavioral patterns.

**Overall Performance** Models exhibit clear performance differences across tiers. Frontier models such as gemini-3-pro-preview and claude-opus-4.5 consistently dominate, achieving top-tier Answer Quality ( $> 4.70$  Completeness) and Success Rate ( $> 88\%$ ). Notably, deepseek-v3.2 emerges as a strong open-weight contender, attaining the highest Recovery Rate (90.6%) and Flexibility (72.4%), even surpassing proprietary models in robustness. In contrast, gpt-4o-mini struggle in open-world settings, posting the lowest Completeness (1.13) and Grounding (0.85), indicating limited readiness for complex autonomous workflows.

In comparison, gpt-5.2 shows signs of reduced stability when coordinating long-horizon tasks. Despite possessing the strongest Grounding (3.80) among other models, it falters significantly in task completion. Qualitative analysis attributes this decline to a pattern of early abandonment: as early as the second turn, the model retreats into passivity, discarding systematic planning and rigorous search. Instead of actively navigating the tool space, it resorts to hallucinating non-existent interfaces to bypass complex sub-tasks. Consequently, critical requirements remain unsatisfied, resulting in a severely penalized Completeness score.

**Behavioral Analysis** Beyond simple rankings, analyzing the sub-metrics reveals critical insights into how models fail or succeed.

Higher tool-using rate does not imply higher successful rate: gpt-4o-mini exhibits the highest volume of Tool Calls (51.71) and # Turns (6.45), yet yields the lowest output quality. This inverse correlation indicates a failure pattern: the agent falls into a “looping” trap, repeatedly invoking tools without effectively synthesizing observations, whereas

gemini-3-pro-preview converts similar high activity (47.86 calls) into superior completeness (4.75).

All LLMs exhibit strong planning ability, but their execution ability is not aligned and cause significant gap in task successful rate. Across the board, Goal Decomposition scores are consistently high (7.7–8.6), proving that most models function as competent Planners. However, the divergence in Tool Calls and Success Rate shows that the bottleneck lies in the Actor’s endurance. For instance, qwen3-235b-a22b plans well (8.51) but executes poorly (11.15 calls, 2.56 completeness), failing to sustain the necessary action sequence.

**Planning vs. Endurance.** Table 3 contrasts initial action against final standing to decouple reasoning from resilience. Despite the strong first-turn score, gpt-5.2 collapses the most over time, dropping five ranks, showing that doing well at the start does not mean the model can sustain performance over longer interactions. Conversely, gemini-3-pro-preview and glm-4.6v overcome mediocre starts to climb +4 and +5 positions respectively, proving that in open-world settings, long-term reflection are more critical than a strong start. Deepseek-v3.2 strikes the best balance, maintaining high quality with zero rank drift and minimal steps.

## 4.3 Human Alignment Measurement

We assess the stability of our LLM evaluation on  $N = 40$  queries (Table 4). Concretely, we ask human annotators independently rank anonymized trajectories, and we compute Spearman correlations for human–human and human–judge agreement. The resulting average human–human agreement is  $\rho = 0.773$ . In comparison, deepseek-v3.2 ( $\rho = 0.759$ ) and gpt-5.1 ( $\rho = 0.733$ ) exhibit close alignment with the human consensus. Although gpt-4o attains a lower mean correlation ( $\rho = 0.688$ ), the uncertainty ranges of human–human and human–LLM correlations substantially overlap when accounting for their respective standard deviations. Together, these results demonstrate the human-level reliability and robustness of the ToolGym scoring framework.

## 4.4 Characteristic Analysis

To further analyze model behavior in constraint-dense environments, we map execution logs to five anthropomorphic dimensions, including (i) *Diligence*, commitment to executing every subgoal through extensive reasoning steps per turn, (ii) *Pru-*

Model	Overall	Quality		Robustness		Constraint		Planning	
		Comp.	Grnd.	Recov. %	Flex. %	Format %	# Calls	Decomp.	
gemini-3-pro-preview	<b>5.87</b>	<b>4.75</b>	2.58	89.0	68.8	53.9	47.9	<b>8.66</b>	
claude-opus-4.5	5.42	4.70	2.93	83.7	60.8	51.0	45.2	7.72	
deepseek-v3.2	4.97	4.00	2.18	<b>90.6</b>	<b>72.4</b>	39.5	21.7	8.04	
glm-4.6v	4.86	4.01	1.18	71.5	57.3	34.2	18.0	8.50	
grok-4	4.78	3.80	1.95	89.0	63.6	<b>68.3</b>	27.4	8.28	
gpt-oss-120b	4.66	3.42	1.28	72.7	59.7	35.8	14.4	8.10	
gpt-5.2	4.43	3.42	<b>3.80</b>	79.3	55.4	12.4	29.2	7.73	
qwen3-235b-a22b	3.53	2.56	1.17	88.1	66.1	31.3	11.2	8.51	
gpt-4o-mini	3.07	1.13	0.85	50.6	39.7	3.3	51.7	7.71	

Table 2: Main Leaderboard Summary. We report the Overall Score alongside key metrics: **Quality** (Completeness & Grounding), **Robustness** (Recovery Rate & Flexibility), **Constraint** (Format), and **Planning** (Avg. Tool Calls & Goal Decomposition). See Appendix E for the full breakdown.

Model	Avg. Steps (Turn 1)	Turn 1 Result	Rank Shift (long horizon)
gpt-5.2	20.22	<b>5.00</b>	↓5
claude-opus-4.5	30.20	4.96	—
grok-4	25.79	4.72	↓2
deepseek-v3.2	<b>10.78</b>	4.67	—
gemini-3-pro	33.19	4.62	↑4
qwen3-235b-a22b	10.75	4.56	↓2
gpt-oss-120b	12.61	4.03	—
glm-4.6v	16.59	3.55	↑5
gpt-4o-mini	22.50	3.04	—

Table 3: First Turn Dynamics vs. Long-horizon Outcome. **Turn 1 Performance** (0–10) indicates initial planning quality. **Rank Shift** tracks how the model’s standing changes when moving from the first turn to the final answer completeness. ↑ indicates models that gain competitive advantage over long horizons; ↓ indicates models that lose ground due to instability.

Left	vs.	Right	$\rho$ (avg. $\pm$ std.)
Human	vs.	Human	$0.773 \pm 0.075$
deepseek-v3.2	vs.	Human	$0.759 \pm 0.073$
gpt-4o	vs.	Human	$0.688 \pm 0.109$
gpt-5.1	vs.	Human	$0.733 \pm 0.091$

Table 4: Comparison of human-human agreement and human-LLM alignment measured by Spearman  $\rho$  on the same set of 40 queries.

*dence*, strict reliance on external verification, measured by the rarity of triggering the “no-tool-for-3-turns” termination condition, (iii) *Grit*, resilience in adversity, defined as the combination of robustness and flexibility, (iv) *Introspection*, persistence in extending dialogue depth to chase down missing details rather than giving up, and v *Strategic*, preference for explicit task decomposition.

During our open-world evaluations, we observe distinct decision-making and planning styles across models. For instance, gemini-3-pro-preview be-

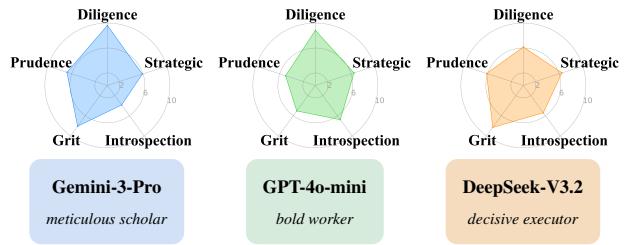


Figure 2: Five-dimensional personality radar charts of different MCP-based agents.

haves like a *meticulous scholar*, choosing to expend extensive acting-reasoning steps to rigorously verify every subgoal before moving forward, whereas gpt-4o-mini is more *impulsive*, often rushing into tool loops without grounding its actions, triggering stagnation checks. Figure 2 presents representative models with their behavioral style.

## 5 LLM Training with ToolGym

### 5.1 Data Curation

We leverage TOOLGYM not just for evaluation, but to curate a high-efficiency instruction tuning dataset for the Actor agent. To target where agents must autonomously translate abstract user intent into concrete tool search queries, we curate trajectories from the 50 seed tasks defined in our experimental setup. We isolate the most cognitively demanding phase by strictly filtering for valid first-round actions generated by all candidate models where translates task intentions into actionable steps. This selection process generates complex logic into a compact dataset of 1,170 samples, ensuring the model learns the core reasoning of tool selection rather than simple pattern matching.

Table 5: Comparison of Pass Rates on BFCL and MCP-Universe benchmarks. TOOLGYM achieves best-in-class performance using significantly less training data.

Model	Data Size	BFCL	MCP-Uni.
<i>Qwen2.5-7B</i>			
+ Toucan	119k	27.18%	15.28%
+ ToolACE	11.3k	27.06%	2.23%
<b>+ ToolGym</b>	<b>1.2k</b>	<b>28.58%</b>	<b>15.30%</b>
<i>Qwen3-8B</i>			
+ Toucan	119k	27.39%	6.67%
+ ToolACE	11.3k	29.49%	3.29%
<b>+ ToolGym</b>	<b>1.2k</b>	<b>30.05%</b>	<b>8.86%</b>

## 5.2 Results Analysis

Table 5 highlights the superior efficiency and generalization of TOOLGYM across both Qwen2.5-7B and Qwen3-8B architectures. On the BFCL benchmark, models fine-tuned on TOOLGYM achieve top-tier accuracy (up to 30.05%), surpassing baselines trained on significantly larger datasets. This suggests that training on complex, constrained trajectories forces the model to pay closer attention to schema definitions and tool selections, thereby enhancing precision even on static tasks. On MCP-Universe, which evaluates generalization to real-world tools, our model maintains robust performance on both model while baselines suffer degradation. This confirms a strong generalization effect: the reasoning skills acquired from TOOLGYM’s ambiguous scenarios enable the agent to effectively execute task in real-world environments. Overall, TOOLGYM demonstrates extreme data efficiency: achieving robustness using only 1,170 samples, far less data samples than competing methods.

## 6 Related Work

**LLM-based Agents.** Research on autonomous agents has evolved to address three fundamental limitations of static LLMs. The first challenge was *bridging internal reasoning with external execution*. While early methods like CoT (Wei et al., 2022) enhanced reasoning, they remained closed systems. This was resolved by the ReAct framework (Yao et al., 2022), which introduced iterative reasoning-action loops to enable interaction with real-world APIs. The second challenge involved *escaping restricted action spaces*. To move beyond rigid schema-based function calling, agents adopted Code-as-Action (Wang et al., 2023, 2024) for open-ended logic generation, or computer-using agents (Zhou et al., 2024; Xie et al., 2024) that per-

ceive and manipulate raw visual interfaces directly. Finally, to tackle *long-horizon complexity and error accumulation*, systems shifted focus toward self-correction and distributed labor. Mechanisms like Reflexion (Shinn et al., 2023) and SwiftSage (Lin et al., 2023) decoupled planning from execution to enhance resilience, while Multi-Agent Systems (Li et al., 2023a; Hong et al., 2023) introduced role specialization to decompose intricate workflows.

**Tool-Using Benchmarks.** The evaluation of tool-using agents has evolved from atomic function calling to complex, open-ended problem-solving. Early benchmarks like API-Bank (Li et al., 2023b) and APIBench (Patil et al., 2024) focused on single-turn selection and schema compliance. To assess reasoning capabilities, ToolBench (Qin et al., 2024) introduced multi-step chaining, while GAIA (Mialon et al., 2023b) and AgentBench (Liu et al., 2024b) targeted general agency in long-horizon objectives. However, these evaluations largely operate within *static or closed contexts*, where agents are either provided with a pre-assigned "ground truth" toolset (e.g., BFCL (Patil et al.)) or confined to sanitized sandboxes (e.g., The Agent Company (Xu et al., 2024)). Consequently, the field is now shifting toward *active retrieval* and *open-world selection*. Works like MCP-ZERO (Fei et al., 2025) and LIVEMCPBENCH (Mo et al., 2025) require agents to autonomously query registries for capabilities, marking a fundamental transition from using to selecting and reasoning.

## 7 Conclusion

This paper presented an open-world tool-using environment designed to realistically and scalably evaluate and improve tool-using LLM agents under large tool pools, long-horizon workflows, complex constraints, and unreliable tool states. Built on 5,571 curated tools across 276 common apps where 206 of them are common apps, our environment combines a task creation engine for multi-tool, wild-constraint workflows with a flexible state controller that injects failures to stress-test robustness, and we further introduced a planner–actor agent framework to disentangle deliberate reasoning from step-wise execution. Our experiments reveal key gaps in current models, including the misalignment between planning and execution abilities, constraint following as the dominant bottleneck, and highlighting DeepSeek-v3.2 as the most robust model under disruptions. Moreover, trajectories

collected automatically from our environment enable highly data-efficient training: fine-tuning with only 1,170 samples outperforms baselines trained on 119k samples. All the above results demonstrate the environment’s value as both a rigorous benchmark and an effective data engine.

## 8 Limitations

While ToolGym provides a scalable and realistic environment for training and testing tool-using agents under large tool pools, our current study has several limitations. Our evaluation set is limited, as we synthesize 50 scenarios, which is not enough to cover the large number of possible tool–server–constraint combinations. Therefore, the leaderboard we report may not reflect performance on less-covered domains, rarely used tools, or less common constraint combinations. In addition, we mainly evaluate frontier and large open models, and we do not systematically test smaller models (e.g., models with <10B parameters). As a result, our conclusions about quality, robustness, and constraint following may not hold for smaller models.

## References

- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *ACL (Findings)*.
- Xiang Fei, Xiawu Zheng, and Hao Feng. 2025. Mcp-zero: Proactive toolchain construction for llm agents from scratch. *arXiv preprint arXiv:2506.01056*.
- Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. 2024. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. In *ACL (Findings)*.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.
- Jeff Johnson, Matthijs Douze, and Herve Jegou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(03):535–547.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. 2023a. Camel: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.
- Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, and 1 others. 2025. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. *arXiv preprint arXiv:2510.25726*.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023b. Api-bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116.
- Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2023. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36:23813–23825.
- Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong WANG, and 1 others. 2024a. Toolace: Winning the points of llm function calling. In *The Thirteenth International Conference on Learning Representations*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuna Yang, and 1 others. 2024b. Agentbench: Evaluating llms as agents. In *ICLR*.
- Zhiwei Liu, Jielin Qiu, Shiyu Wang, Jianguo Zhang, Zuxin Liu, Roshan Ram, Haolin Chen, Weiran Yao, Shelby Heinecke, Silvio Savarese, and 1 others. 2025. Mcpeval: Automatic mcp-based deep evaluation for ai agent models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 373–402.
- Ziyang Luo, Zhiqi Shen, Wenzhuo Yang, Zirui Zhao, Prathyusha Jwalapuram, Amrita Saha, Doyen Sahoo, Silvio Savarese, Caiming Xiong, and Junnan Li. 2025. Mcp-universe: Benchmarking large language models with real-world model context protocol servers. In *Workshop on Scaling Environments for Agents*.

- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, and 1 others. 2023a. Augmented language models: a survey. *Transactions on Machine Learning Research*.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023b. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- Guozhao Mo, Wenliang Zhong, Jiawei Chen, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2025. Livemcpbench: Can agents navigate an ocean of mcp tools? *arXiv preprint arXiv:2508.01780*.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.
- Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37:126544–126565.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toollm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. In *Intrinsically-Motivated and Open-Ended Learning Workshop @ NeurIPS2023*.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better llm agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, Yujia Bao, and 1 others. 2025. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. In *Workshop on Scaling Environments for Agents*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.*
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.
- Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, and 1 others. 2024. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*.
- Zhangchen Xu, Adriana Meza Soria, Shawn Tan, Anurag Roy, Ashish Sunil Agrawal, Radha Poovendran, and Rameswar Panda. 2025. Toucan: Synthesizing 1.5 m tool-agentic data from real-world mcp environments. *arXiv preprint arXiv:2510.01179*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.
- Junjie Ye, Guanyu Li, Songyang Gao, Caishuang Huang, Yilong Wu, Sixian Li, Xiaoran Fan, Shihan Dou, Tao Ji, Qi Zhang, and 1 others. 2025. Tooleyes: Fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. In *COLING*.
- Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2024. Webarena: A realistic web environment for building autonomous agents. In *12th International Conference on Learning Representations, ICLR 2024*.

## A Evaluation Metrics

To strictly evaluate open-world orchestration, we measure performance across four dimensions:

**Quality.** (i) *Completeness* quantifies the resolution of user intent to ensure the task is actually solved, measured by an LLM comparing the final response against the task requirements. (ii) *Grounding* ensures factual reliability by an LLM verifier, checking that the answer is strictly derived from tool observation logs and penalizing any hallucinations not supported by the execution history.

**Robustness.** (i) *Schema Compliance* ensures fundamental functionality by validating every tool call against the official JSON schema definitions to reject syntax or type errors. (ii) *Recovery Rate* measures self-correction by calculating the conditional probability of a successful execution in turn  $t + 1$  given a failure in turn  $t$ . (iii) *Flexibility* evaluates strategic adaptability by calculating the conditional probability that the agent takes an alternative execution plan after a State Controller-injected error, thereby identifying agents capable of adaptive re-planning under failure.

**Constraint.** We evaluate adherence against the structured constraints pre-defined in the seed query: (i) *Order & Diversity* ensure logical sequencing by comparing execution timestamps and server counts against the query’s explicit dependency graph. (ii) *Format* guarantees structural compliance by using an LLM to verify that both the output structure and content match the patterns specified in the task requirements. (iii) *Tradeoff* evaluates reasoning using an LLM to verify if the resolution of conflicting objectives aligns with the implicit preferences embedded in the user request.

**Planning.** (i) *Goal Decomposition* evaluates strategic alignment by employing an LLM judge to score the semantic overlap between the generated plan and the ground-truth reference rationales. (ii) *Progress Tracking* assesses state estimation accuracy by using an external evaluator to cross-check the Planner’s status labels (*e.g.*, ‘Complete’) against the actual tool execution logs. (iii) *Efficiency* quantifies operational conciseness by recording the raw number of interaction turns required to successfully resolve the query.

We employ a unified scoring protocol: deterministic metrics (*e.g.*, Schema, Order) are reported as percentages(%) over all evaluation instances, while

semantic assessments (*e.g.*, Completeness, Trade-off) are graded by the LLM Judge on a 10-point scale.

## B Evaluated Models

We evaluate 9 representative models (Table 2) spanning proprietary and open-weight paradigms to form a holistic benchmark. This includes frontier closed-source models (gpt-5.2, gemini-3-pro-preview, claude-opus-4.5, grok-4), popular open-weight systems (deepseek-v3.2, glm-4.6v, qwen3-235b-a22b, gpt-oss-120b), as well as a cost-optimized baseline (gpt-4o-mini) to examine long-horizon reasoning abilities across mainstream models.

## C Training Implementation

We conduct experiments on Qwen2.5-7B-Instruct and Qwen3-8B to demonstrate the effectiveness and robustness of our dataset across different model architectures and scales. To evaluate the advantages of our dataset, we compare it against Toucan (Xu et al., 2025) and ToolACE (Liu et al., 2024a). All datasets are converted into a data format compatible with the ms-swift training framework.<sup>2</sup> Model training follows the Hermes-style agent supervision paradigm,<sup>3</sup> which explicitly models tool usage through tool invocation and tool response messages. For baseline comparisons, we use the full ToolACE dataset containing 11.3K samples, and additionally incorporate a subset of Toucan data drawn from the SFT class, which comprises 119K trajectories in total.

## D LLM Usage Statement

Large language models (LLMs) were used solely for language editing—improving grammar, clarity, and overall readability. They did not generate or modify the manuscript’s scientific content, conceptual contributions, methodology, or experimental results. The authors assume full responsibility for the final text and its accuracy.

## E Detailed Evaluation Results

<sup>2</sup><https://github.com/modelscope/ms-swift>

<sup>3</sup><https://huggingface.co/NousResearch/Hermes-2-Pro-Llama-3-8B>

Model	Overall	Answer Quality			Tool Use Robustness			Constraint Following			Long-horizon			
		Score	Completeness	Grounding	Success Rate	Recovery Rate	Flexibility	Order	Info Diversity	Format	Tradeoff	Tool Calls	# Turns	Progress Tracking
gemini-3-pro-preview	5.87	4.75	2.58	88.8%	89.0%	68.8%	53.8%	97.8%	53.9%	13.3%	47.86	3.20	7.60	8.66
claude-opus-4.5	5.42	4.70	2.93	92.7%	83.7%	60.8%	65.4%	73.3%	51.0%	33.7%	45.16	4.01	6.41	7.72
deepeek-v3.2	4.97	4.00	2.18	87.5%	90.6%	72.4%	73.1%	70.0%	39.5%	17.9%	21.73	4.92	6.46	8.04
glm-4.6v	4.86	4.01	1.18	84.8%	71.5%	57.3%	75.6%	52.2%	34.2%	11.5%	18.03	3.27	7.20	8.50
grok-4	4.78	3.80	1.95	87.8%	89.0%	63.6%	64.1%	92.2%	68.3%	35.5%	27.37	2.55	6.02	8.28
gpt-oss-120b	4.66	3.42	1.28	86.3%	72.7%	59.7%	87.2%	38.9%	35.8%	13.3%	14.40	3.14	6.53	8.10
gpt-5.2	4.43	3.42	3.80	85.5%	79.3%	55.4%	71.6%	37.2%	12.4%	12.5%	29.20	2.30	5.62	7.73
qwen3-235b-a22b	3.53	2.56	1.17	87.9%	88.1%	66.1%	80.8%	43.3%	31.3%	8.6%	11.15	4.41	6.93	8.51
gpt-4o-mini	3.07	1.13	0.85	87.5%	50.6%	39.7%	85.9%	46.7%	3.3%	0.0%	51.71	6.45	6.00	7.71

Table 6: Full Leaderboard breakdown including sub-metrics for Answer Quality, Tool Use Robustness, Constraint Following, and Long-horizon planning.