M^3-BENCH: MULTI-MODAL, MULTI-HOP, MULTI-THREADED TOOL-USING MLLM AGENT BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

We present M^3-Bench, the first benchmark for evaluating multimodal tool use under the Model Context Protocol. The benchmark targets realistic, multi-hop and multi-threaded workflows that require visual grounding and textual reasoning, cross-tool dependencies, and persistence of intermediate resources across steps. We introduce a similarity-driven alignment that serializes each tool call, embeds signatures with a sentence encoder, and performs similarity-bucketed Hungarian matching to obtain auditable one-to-one correspondences. On top of this alignment, we report interpretable metrics that decouple semantic fidelity from workflow consistency. The benchmark spans 27 servers with 232 tools, and provides standardized trajectories curated through an Executor & Judge pipeline with human verification; an auxiliary four large language models (LLMs) judge ensemble reports end-task Task Completion and information grounding. Evaluations of representative state-of-the-art Multimodal LLMs (MLLMs) reveal persistent gaps in multimodal MCP tool use, particularly in argument fidelity and structure consistency, underscoring the need for methods that jointly reason over images, text, and tool graphs. Our Benchmark's anonymous repository is at `https://anonymous.4open.science/r/M3-Bench-E8848/`

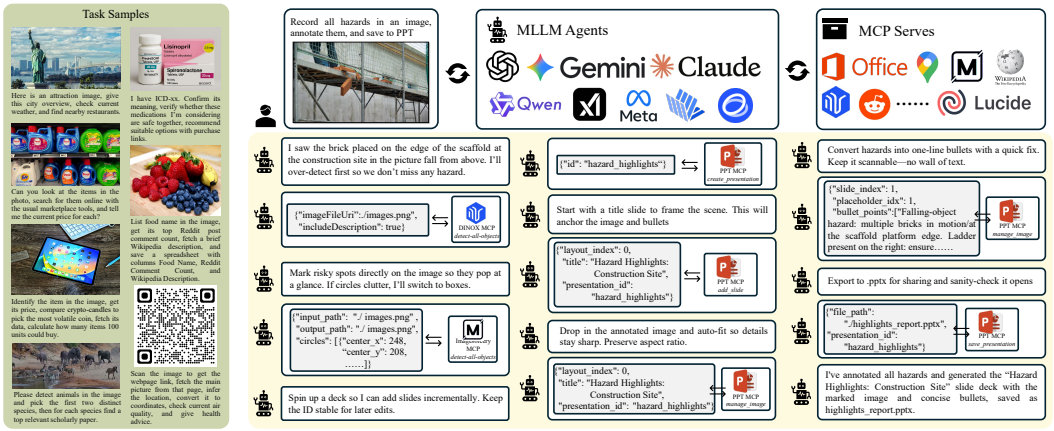
1 INTRODUCTION

We introduce M^3-Bench, the *Multi-Modal*, *Multi-Hop*, *Multi-Threaded* MCP Benchmark, as a principled evaluation suite for multimodal tool use under the Model Context Protocol (MCP). Multimodal Large Language Models (MLLMs) have recently shown that, once they are allowed to perform function calling (tool use), they can query external services and reason over information that is not contained in their parameters (Schick et al., 2023; Yao et al., 2023; Shen et al., 2023; Comanici et al., 2025; Anthropic, 2025). MCP specifies how models communicate with heterogeneous tools through standardized servers, which makes execution more reproducible across systems (Model Context Protocol Working Group, 2025). However, existing MCP benchmarks are mostly text-only and focus on linear API planning or database queries (Patil et al., 2025; Yao et al., 2025; Gao et al., 2025; Liu et al., 2025). A systematic evaluation of *multimodal* MCP workflows, where images and text jointly condition tool calls and results, is still missing (Anthropic, 2024; Luo et al., 2025).

The core challenge in real-world MCP trajectories is visual grounding: multimodal tool invocation hinges on correctly interpreting the image before any tool can be parameterized. In Figure 1a, the agent receives a photograph and must first resolve the landmark/city from visual cues, only then can it condition subsequent MCP calls; the agent must fully recognize the well-stocked and well-displayed products on the shelf in a photograph before proceeding to the next MCP tool call. Second, real-world MCP trajectories are not single-shot calls. They are multi-hop and contain causally dependent operations, and they frequently execute several tool calls in parallel within one step (Mialon et al., 2024; Zhou et al., 2024; Li et al., 2024; Yu et al., 2024). The task in Figure 1b illustrates this setting. The agent receives an image of a hazardous construction scene and a reporting-style instruction.

Inside one step, actions that do not depend on each other, such as adding an image and updating bullets, can be executed concurrently. Across steps, operations that do depend on earlier results, such as annotating before inserting or creating the deck before saving, must follow the ground-truth order. These properties make simple string matching or linear-sequence scoring inadequate (Mialon et al., 2024; Zhou et al., 2024), and it is precisely what our benchmark aims to stress. We define

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: Overview of tasks and a realistic multimodal MCP workflow. (a) Representative task samples in M^3-Bench. (b) A realistic multimodal MCP workflow: the agent (i) calls a vision MCP server to detect and mark hazards in the input image, (ii) calls Imagesorcery MCP to process the image, and (iii) calls an office/productivity MCP server to create and populate a slide deck.

two important concepts: **Multi-Hop** refers to workflows with more than one causally dependent step, where later actions consume artifacts produced earlier (Ye et al., 2025; He et al., 2024); **Multi-Threaded** refers to order-independent tool calls executed within a single step under shared state, allowing safe parallelism while preserving cross-step causality (Definitions 1 and 2 in Section 4.1).

Table 1 contrasts M^3-Bench with prior tool-use evaluations along 8 axes. Critically, *Multi-Hop & Threaded* captures cross-step causal dependencies together with safe within-step parallelism; *Multimodality* requires joint image–text conditioning; and *Similarity Metric* denotes an LLM-free, similarity-aware alignment of predicted to reference calls (see Section 3). Most existing benchmarks are text-only and/or lack explicit multi-threaded causality, and none pair multimodality with a similarity-aware alignment; M^3-Bench satisfies all three while retaining the key advantages of prior work.

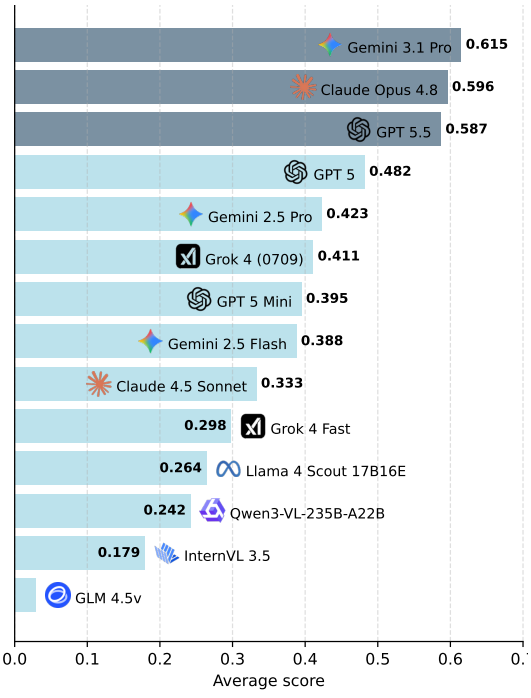


Figure 2: Summary leaderboard of state-of-the-art (SOTA) MLLMs on M^3-Bench (higher is better).

We align predicted and reference tool calls with a *similarity-bucketed Hungarian alignment*: each call is serialized, embedded with a fixed sentence encoder, and one-to-one matched within tool-name buckets under weak/strong cosine thresholds. This gives deterministic, auditable correspondences without relying on an LLM judge for call-level scoring. On top of this alignment, we report a compact, recall-aware suite that separates semantic fidelity from workflow consistency. A small four-model judge ensemble is used for evaluating the overall quality of the trajectory to retain some of the advantages of the LLM judge feature. Besides, to standardize references, our experiments also provide a *best trajectory* obtained via an Executor–Judge loop.

Contributions. **(1)** We present, to our knowledge, the first benchmark explicitly targeting **multimodal** MCP workflows. **(2)** Our repository provides an end-to-end pipeline for generating MCP best trajectories, with an optional lightweight human verification pass to enhance stability. **(3)** We introduce a structure-aware metric suite that aligns tool calls via bucketed Hungarian matching on high-dimensional feature similarities, yielding robust and auditable correspondences for effectively

diagnosing distinct error modes. **(4)** We evaluate a wide range of mainstream MLLMs and conduct a comprehensive assessment across tasks, budgets, error types, and other factors.

2 Related Work

LLM Tool Use and Protocols Early studies showed that large language models (LLMs) can autonomously decide *when/what/how* to call tools via APIs (Schick et al., 2023; Mialon et al., 2024; Froger et al., 2025; Qin et al., 2023; Mo et al., 2025; Jin et al., 2025a;b;c; Dong et al., 2026). Reliability improves with API-aware finetuning and retrieval (Patil et al., 2024); instruction/data scaling expands tool repertoires and evaluations (ToolBench) (Qin et al., 2023); and hierarchical retrieval with self-reflection addresses large tool spaces and brittle protocols (AnyTool/AnyToolBench) (Du et al., 2024). Benchmarks targeting the *decision to use a tool* and tool selection (MetaTool Benchmark) complement these directions (Huang et al., 2024). To reduce integration friction across tools, the Model Context Protocol (MCP) standardizes interfaces (Anthropic, 2024; Model Context Protocol Working Group, 2025).

Benchmark	Domains	Tools	MCP Ecosystem	Info Grounding	Fuzzy Task	Multi-Hop & Thrd	Multimodality Sim. Metric
ToolBench	49	3451	✗	✗	✗	✗	✗ ✗
BFCL v3	8	24	✗	✗	✗	✗	✗ ✗
τ-Bench	2	28	✗	✗	✗	✗	✗ ✗
MCP-RADER	9	42	✓	✗	✗	✗	✗ ✗
MCPEval	5	19	✓	✗	✗	✗	✗ ✗
MCPToolBench++	6	87	✓	✓	✗	✗	✗ ✗
MCP-Atlas	40+	300+	✓	✓	✗	✗	✗ ✗
MCP-Universe	6	133	✓	✓	✓	✗	✗ ✗
MCP-Bench	28	250	✓	✓	✓	✓	✗ ✗
M^3-Bench (Ours)	**27**	**232**	✓	✓	✓	✓	✓ ✓

Table 1: Comparisons to existing tool-using benchmarks.

Benchmarks for Tool Use and MCP Beyond general agent evaluations such as AgentBench (Liu et al., 2024) and web-agent suites like WebArena and BrowserGym (Zhou et al., 2024; Le Sellier De Chezelles et al., 2024), MCP-native efforts emphasize multi-hop use over live servers: *MCP-Bench* links LLMs to dozens of MCP servers and hundreds of tools (Wang et al., 2025b); *MCP-Universe* provides a framework grounded in real-world servers and long-horizon tasks (Salesforce AI Research, 2025a;b); *MCPToolBench++* scales tool categories and tasks via marketplace-sourced servers (Fan et al., 2025); and *MCP-Atlas* curates a human-authored dataset and leaderboard for multi-tool workflows (Scale AI Research, 2025; Scale AI, 2025). Relative to these, we target *multimodal* tasks and introduce a similarity-aware alignment with structure-sensitive diagnostics.

3 M^3-Bench

3.1 Benchmark Construction

We generate benchmark trajectories with an Executor–Judge loop that decouples tool execution from adjudication (Figure 3). For each image–task pair, a human instruction fixes intent and permissible tools. Given the prompt, image, and registry, the Executor produces a stepwise trajectory; the Judge audits and consolidates it into a standardized *optimal trajectory* for alignment. All prompts are in Appendix B.2. Via system prompts, the Executor receives the full tool list (with docstrings) and outputs the planned tool set for the current step—*selection only* to stabilize reasoning and keep the chain controllable. (See Appendix Prompt: Plan.)

Process The Executor returns strictly valid JSON `tool_calls` (each with `name` and `arguments`). A multi-hop workflow allows multi-threaded interaction with the MCP Ecosystem within a single step. Structured outputs are appended to the dialogue, and any file-image paths from MCP tools persist in a temporary workspace for later steps. (See Appendix Prompt: Process.)

Decision and Final Reply After each step, a binary prompt checks whether evidence suffices to conclude; if so, the Executor emits only the final answer. (See Appendix Prompt: Decision, Final.)

Judge and Summarize The Judge consumes the initial prompt, image, and full Executor tool-call chain. Trajectories with causal misordering or factual errors are rejected; otherwise, the Judge reconstructs the *optimal trajectory*, removing failed/redundant calls and merging operations that could run concurrently. Human verifiers audit scope and logical consistency; the audited result becomes the benchmark reference. (See Appendix Prompt: Judge, Prompt: Summarize.)

3.2 MCP ECOSYSTEM

The benchmark contains 211 multimodal MCP cases across 27 scenario domains; detailed goals and input–output constraints are provided in Appendix B, and representative full trajectories appear in Appendix D. Their reference trajectories comprise 644 steps and 1337 MCP tool calls.

To keep trajectories controllable and reproducible, we curate a compact set of MCP servers and tools with minimal functional overlap. We favor orthogonal, well-documented, and stable utilities, discarding redundant or unstable options to reduce confounds during evaluation. Our reference trajectories exercise all 27 MCP servers and 69 of their 232 tools (Figure 4). This diverse ecosystem covers Academic & Knowledge, Science & Space, Geography & Travel, E-commerce & Finance, Office Automation, Weather & Air Quality, and Health, enabling broad capability coverage in M^3-Bench.

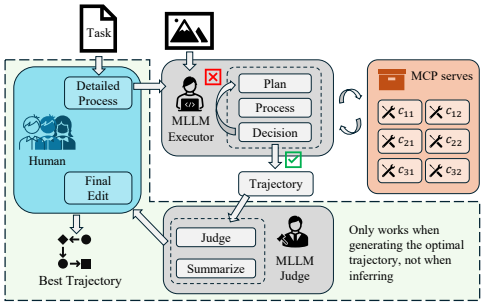


Figure 3: Executor–Judge trajectory pipeline. The Executor plans tools, executes parallel calls within each step, and summarizes; the Judge reconstructs and verifies an optimal trajectory, where c_{ℓ,n_ℓ} denotes the n_ℓ-th tool call in step ℓ.

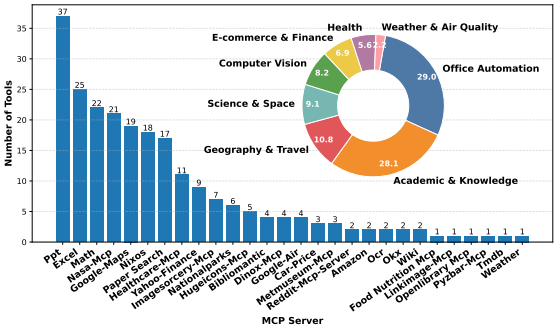


Figure 4: Overview of MCP server ecosystem used in the M^3-Bench. Category distribution of MCP servers; and tool distribution across servers. The inventory follows the MCP design and public specification.

4 METRICS

In Section 4.1, we first establish the mathematical definitions and foundations for our metrics. Building on these theoretical bases, we then construct the Detection Metrics Section 4.2, Trajectory Alignment Metrics Section 4.3, and LLM Judges Metrics Section 4.4.

4.1 MATHEMATICAL FOUNDATIONS

A trajectory is a DAG ordered by turns (*steps*). Denote L as the number of steps in a trajectory $\mathcal{T}$. The ground-truth $\mathcal{T}^\star$ and predicted trajectories $\widehat{\mathcal{T}}$ are: $\mathcal{T}^\star = \{S_1^\star, \ldots, S_{L^\star}^\star\}$, $\widehat{\mathcal{T}} = \{\widehat{S}_1, \ldots, \widehat{S}_{\widehat{L}}\}$, where each step S enumerates its calls c, $S_\ell^\star = \{\, c_{\ell,1}^\star, \ldots, c_{\ell,n_\ell^\star}^\star \,\}$, $\widehat{S}_{\hat{\ell}} = \{\, \hat{c}_{\hat{\ell},1}, \ldots, \hat{c}_{\hat{\ell},n_{\hat{\ell}}} \,\}$. where ℓ is the step of the trajectory, so $n_\ell^\star$ and $n_{\hat{\ell}}$ are per-step call counts. A call is $c = (\texttt{tool}(c), \texttt{args}(c))$, where $\texttt{tool}(\cdot)$ is the categorical tool identifier and $\texttt{args}(\cdot)$ serializes all deterministic inputs.

Definition 1 (Multi-Hop)**.** Let $\mathcal{T} = \{S_1, \ldots, S_L\}$ be a tool-use trajectory with steps $S_\ell = \{c_{\ell,1}, \ldots, c_{\ell,n_\ell}\}$ and inter-call dependency edges across steps. We call $\mathcal{T}$ *multi-hop* if $L \geq 2$ and there exists a dependency $(c_{i,a} \to c_{j,b})$ with $i < j$, where $a, b \in \mathbb{N}$ and $a \in [1, n_i]$, $b \in [1, n_j]$.

Definition 2 (Multi-Threaded)**.** A step S_ℓ is *multi-threaded* if it contains at least two calls $c_{\ell,a} \neq c_{\ell,b}$ that are order-independent given the shared state at depth ℓ (no dependency edge between them).

These definitions make explicit that a trajectory is a structured object with cross-step causal edges and within-step parallelism. A general option is to delegate this alignment to an LLM judge, but it introduces (i) *lineage bias*, where judges trained like the systems under test tend to over-credit them; (ii) susceptibility to *prompt gaming*, as long rationales or CoT responses can inflate scores even when the underlying tool use is wrong; (iii) *weak argument verification*, since judges rarely check whether argument objects are semantically correct or whether order-invariant sets are handled; (iv) *knowledge substitution*, where missing tool evidence is replaced by the model's own world knowledge; and (v)

opacity, as scores lack auditable, one-to-one correspondences between predicted and reference calls. Real MCP trajectories exacerbate these issues: steps often contain *unordered* sets of calls, later calls can *self-correct* earlier errors, and partial success is common. Consequently, the alignment primitive must be deterministic, auditable, and robust to surface-form variation.

Similarity-Bucketed Hungarian Alignment To address this problem, we introduce a similarity computation method based on a text encoder. Each call is flattened to a canonical text string by a serialization function $\phi(\cdot)$ that concatenates the tool name with its arguments serialized via `json.dumps(sort_keys=True)`, which recursively sorts dictionary keys at every nesting level, guaranteeing a deterministic, order-invariant representation for any valid JSON argument structure. We embed serialized calls with a sentence encoder $\mathcal{E}$ and compute cosine similarity: $S_{ij} = \frac{\langle \mathbf{x}_i^\star, \widehat{\mathbf{x}}_j \rangle}{\|\mathbf{x}_i^\star\|_2 \, \|\widehat{\mathbf{x}}_j\|_2} \in [-1, 1]$, where $\mathbf{x}_i^\star = \mathcal{E}\big(\phi(c_i^\star)\big)$ and $\widehat{\mathbf{x}}_j = \mathcal{E}\big(\phi(\hat{c}_j)\big)$. We align with a tool-specific, step-agnostic assignment that yields auditable one-to-one correspondences. Concretely, we serialize every call into a canonical text signature and embed it with the sentence encoder SentenceTransformers (Reimers & Gurevych, 2019) from Textual Encoding and Similarity, giving a similarity matrix S. For each tool k, we form the bucketed index sets $\mathcal{I}_k = \{\, i \mid \texttt{tool}(c_i^\star) = k \,\}$, $\widehat{\mathcal{J}}_k = \{\, j \mid \texttt{tool}(\hat{c}_j) = k \,\}$, so that only same-tool pairs are considered. We convert similarity to cost with a weak screening threshold τ_{weak} to forbid low-similarity edges: $C_{rc}^{(k)} = \begin{cases} 1 - S_{rc}^{(k)}, & S_{rc}^{(k)} \geq \tau_{\text{weak}}, \\ \lambda_{\text{pen}}, & S_{rc}^{(k)} < \tau_{\text{weak}}, \end{cases}$ where $\lambda_{\text{pen}} = 10^6$ is large enough to ensure that below-threshold pairs are never selected. We solve a linear assignment in each bucket via the Hungarian (Kuhn–Munkres) algorithm $\min_{\Pi^{(k)} \in \mathcal{P}} \langle C^{(k)}, \Pi^{(k)} \rangle$ retaining only feasible pairs above the weak threshold to form the global match set: $M = \{(i, j, S_{ij}) \mid S_{ij} \geq \tau_{\text{weak}}\}$.

Properties and rationale. Tool bucketing prevents cross-tool credit; step agnosticism allows *early-wrong, later-correct* fixes to match without forcing same-step alignment; and similarity gating recognizes paraphrase and order-invariant arguments without inflating coverage. Together, these choices deliver: (i) deterministic, auditable correspondences in place of LLM judge; (ii) invariance to within-step call permutations (sets are unordered); and (iii) faithful scoring under *partial success*, misaligned or missing calls simply remain unmatched, while correctly recovered calls still receive credit. This alignment thus targets exactly the failure modes we observe in MCP trajectories while satisfying the three requirements in Section 4.1.

4.2 Detection Metrics

Based on the above mathematical foundations, we introduce detection-based metrics (see Appendix 8 for an overview): **(1)** Systems may call incorrect tools or omit required ones, motivating **Recall**; **(2)** MLLM may make unnecessary calls, resulting in additional waste of resources **Precision**.

Recall and Precision. (Appendix 8 (a)-(b)) These metrics quantify, respectively, how many ground-truth calls are successfully matched (coverage) and how many predicted calls are valid (false-positive control). Given per-sample match sets M_m with counts $N_{\text{gt},m}$ and $N_{\text{pred},m}$, we define $\text{Recall} = \frac{\sum_m |M_m|}{\sum_m N_{\text{gt},m}}$, $\text{Prec} = \frac{\sum_m |M_m|}{\sum_m N_{\text{pred},m}}$, where $|M_m|$ counts matched pairs for sample m, $N_{\text{gt}} = \sum_\ell n_\ell^\star$ and $N_{\text{pred}} = \sum_{\hat{\ell}} n_{\hat{\ell}}$ be the total numbers of calls in $\mathcal{T}^\star$ and $\widehat{\mathcal{T}}$, respectively.

4.3 Trajectory Alignment Metrics

Next, we need to quantify the degree of Trajectory Alignment: **(1)** MLLM may select the correct tool but supply incorrect arguments, requiring an **Average Similarity** measure over serialized calls; **(2)** MLLM may unnecessarily split operations that could run in a single step in a **multi-threaded** scenario, captured by **Step Coherence**; **(3)** MLLM may collapse causally distinct multi-hop subgoals into one step in a **multi-hop** scenario, quantified by **Merge Purity**; **(4)** MLLM may invert the execution order of causally dependent steps in a **multi-hop** scenario, detected by **Order Consistency**.

Model	Detection		Trajectory Alignment				LLM Judge		Avg
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	Score
GLM 4.5v	0.013	0.023	0.011	0.012	0.000	0.013	0.285	0.130	0.029
Qwen2.5-VL-72B	0.143	0.073	0.125	0.125	0.022	0.112	0.303	0.388	0.141
InternVL 3.5	0.224	0.143	0.193	0.191	0.048	0.161	0.325	0.298	0.179
Gemini 2.5 Flash Lite	0.200	0.173	0.168	0.179	0.030	0.152	0.299	0.241	0.180
Claude 4.5 Haiku	0.193	0.480	0.166	0.190	0.053	0.108	0.412	0.246	0.205
Qwen3-VL-235B-A22B	0.260	0.217	0.248	0.236	0.095	0.225	0.336	0.416	0.242
GPT-5 Nano	0.287	0.103	0.276	0.228	0.045	0.258	0.551	0.529	0.247
L4-Scout-17B16E	0.367	0.176	0.325	0.286	0.068	0.290	0.367	0.335	0.264
Grok-4 Fast	0.343	0.359	0.309	0.284	0.122	0.289	0.501	0.378	0.298
Claude 4.5 Sonnet	0.395	0.361	0.354	0.389	0.127	0.257	0.535	0.451	0.333
Gemini 2.5 Flash	0.454	0.397	0.409	0.357	0.211	0.412	0.530	0.474	0.388
GPT-5 Mini	0.481	0.238	0.462	0.392	0.199	0.360	0.652	0.633	0.395
Grok 4 (0709)	0.465	0.387	0.408	0.415	0.177	0.390	0.664	0.638	0.411
Gemini 2.5 Pro	0.501	0.449	0.433	0.464	0.249	0.424	0.520	0.438	0.423
GPT-5	0.627	0.264	0.583	0.502	0.290	0.453	0.671	0.657	0.482
GPT-5.5	**0.745**	0.422	**0.667**	0.533	0.443	0.637	0.574	**0.663**	0.587
Claude Opus 4.8	0.686	**0.668**	0.607	0.507	0.417	0.626	**0.682**	0.659	0.596
Gemini 3.1 Pro	0.734	0.567	0.647	**0.582**	**0.453**	**0.659**	0.603	0.662	**0.615**

Table 2: Cross-model evaluation of multimodal MCP trajectories. Avg Score corresponds to Figure 2. *Note:* Bold/underline = first/second; L4-Scout-17B16E = Llama-4-Scout-17B16E. Sources: Table 17.

Argument Similarity. We summarize the semantic faithfulness of matched arguments as

$$\mathrm{ArgSim} = \frac{\sum_{(i,j)\in M,\, S_{ij}\geq\tau_{\mathrm{strong}}} S_{ij}}{|\{(i,j)\in M : S_{ij}\geq\tau_{\mathrm{strong}}\}|}, \tag{1}$$

where τ_{strong} controls alignment confidence.

Step Coherence. (Appendix 8 (c)) This metric penalizes scattering matches from one ground-truth step across multiple predicted steps, reflecting unnecessary splits of operations that could have been executed within a single **multi-threaded** step. For any call c, let $\mathrm{gt}_{\mathrm{st}}(c)$ and $\mathrm{pred}_{\mathrm{st}}(c)$ denote its step indices in the ground-truth and predicted trajectories. For each ground-truth step ℓ, collect the predicted step indices touched by its matches: $\mathcal{P}(\ell) = \{\ \mathrm{pred}_{\mathrm{st}}(\hat{c}_j)\ \mid\ (i,j)\in M, \mathrm{gt}_{\mathrm{st}}(c_i^\star) = \ell\ \}$.Then we define the per-step coherence and a similarity-weighted global average as

$$\mathrm{sc}(\ell) = \begin{cases} 1, & |\mathcal{P}(\ell)|\leq 1, \\ \dfrac{1}{|\mathcal{P}(\ell)|}, & \text{otherwise}, \end{cases}, \quad \mathrm{StepCoh} = \frac{\sum_\ell w_\ell\, \mathrm{sc}(\ell)}{\sum_\ell w_\ell}, \tag{2}$$

where $w_\ell = \sum_{(i,j)\in M}[\mathrm{gt}_{\mathrm{st}}(c_i^\star) = \ell]$ counts matches originating from ground-truth step ℓ.

Merge Purity. (Appendix 8 (d)) This metric is used to evaluate whether, in a **multi-hop** scenario, different causally related steps are incorrectly merged. For total alignment weight between ground-truth step a and predicted step b: $W_{ab} = \sum_{(i,j)\in M}[\mathrm{gt}_{\mathrm{st}}(c_i^\star) = a]\ [\mathrm{pred}_{\mathrm{st}}(\hat{c}_j) = b]\ S_{ij}$, where $[\cdot]$ denotes the Iverson bracket (indicator function), equal to (1) when its condition is true and (0) otherwise. Define column sums $S_b = \sum_a W_{ab}$, total mass $S = \sum_b S_b$, and column prior $P(b) = S_b/S$. The conditional distribution of ground-truth steps given a predicted step is $q_{a|b} = W_{ab}/S_b$ for active columns with $S_b > 0$. The normalized purity is:

$$\mathrm{MergePur} = 1 - \frac{H(G\mid P)}{\log G_{\mathrm{act}}}, \quad H(G\mid P) = \sum_{b:S_b>0} P(b)\Big(-\sum_{a:W_{ab}>0} q_{a|b}\log q_{a|b}\Big), \tag{3}$$

where $G_{\mathrm{act}} = |\{a : \sum_b W_{ab} > 0\}|$ counts active ground-truth steps, $H(\cdot)$ uses natural logarithm, and larger values indicate predicted steps concentrating mass on single ground-truth steps (i.e., less over-merging); G denotes the random variable of ground-truth step indices and P denotes the random variable of predicted step indices; $G_{\mathrm{act}} = |\{a : \sum_b W_{ab} > 0\}|$ is the number of active ground-truth steps. The normalization by $\log G_{\mathrm{act}}$ maps $[0, \log G_{\mathrm{act}}]$ to $[0, 1]$.

Order Consistency. (Appendix 8 (e)) This metric is used to evaluate whether, in a **multi-hop** scenario, the order of causally related calls is incorrectly swapped, capturing order consistency across steps. Let each match $(i, j) \in M$ be represented by its step pair $(a, b) = (\mathrm{gt}_{\mathrm{st}}(c_i^\star), \mathrm{pred}_{\mathrm{st}}(\hat{c}_j))$. Over all unordered pairs of matches that are comparable in both axes, we compute:

$$\mathrm{OrdCons} = 1 - \frac{1}{Q} \sum_{(u,v)} \big[(a_u - a_v)(b_u - b_v) < 0\big], \tag{4}$$

where the sum runs over pairs with $a_u \neq a_v$ and $b_u \neq b_v$, Q is the number of such pairs, and the indicator records inversions between the ground-truth and predicted step orders; higher values indicate fewer inversions.

To penalize degenerate cases with very few matches (weak systems that match only a few easy MCP tools from inflating structure-aware metrics, leading to a very small denominator in the metric calculation, and thus an inflated outcome), we weight per-sample metrics by their recall. Let the per-sample recall be $r_m = \frac{|M_m|}{N_{\mathrm{gt},m}}$. For any per-sample metric $F_m \in \{\mathrm{ArgSim}, \mathrm{StepCoh}, \mathrm{MergePur}, \mathrm{OrdCons}\}$, we report a recall-covered score: $F^{\mathrm{cov}} = \frac{\sum_m N_{\mathrm{gt},m}\, r_m\, F_m}{\sum_m N_{\mathrm{gt},m}}$.

Edge-case conventions. When the denominator of a metric is zero we adopt the following defaults, consistent with the implementation: **(i)** $\mathrm{ArgSim} = 0$ when no match exceeds τ_{strong}; **(ii)** $\mathrm{StepCoh} = 0$ when $|M| = 0$ (no matched calls); **(iii)** $\mathrm{MergePur} = 1$ when $G_{\mathrm{act}} \leq 1$ (single or no active ground-truth step); **(iv)** $\mathrm{OrdCons} = 1$ when $Q = 0$ (fewer than two comparable matched step pairs). By construction $F^{\mathrm{cov}} \leq \mathrm{Recall}$; per-task tables may show differences of ± 0.001 due to floating-point rounding in the per-step weighted aggregation.

4.4 LLMs Judges Metrics

To leverage the advantages of large language models (LLMs) in complex logical reasoning, we introduce a metric based on LLMs Judges to assess: **(1)** Whether the MLLM achieves **Task Completion**; **(2)** Whether the answer provided by the MLLM possesses MCP **Information Grounding**.

Task Completion. For each instance m, a fixed set of four LLM judges $\mathcal{J}$ produces scalar scores $\{s_m^{(j)}\}_{j \in \mathcal{J}} \subset [0, 1]$. We sort the four scores in non-decreasing order $s_m^{(1)} \leq s_m^{(2)} \leq s_m^{(3)} \leq s_m^{(4)}$ and compute a trimmed mean by averaging the two interior values (Prompt: Appendix B.2):

$$\mathrm{TaskComp}_m = \tfrac{1}{2} \sum_{j \in \mathcal{J} \setminus \{j_{\min},\, j_{\max}\}} s_m^{(j)}, \quad j_{\min} = \arg\min_{j \in \mathcal{J}} s_m^{(j)},\ j_{\max} = \arg\max_{j \in \mathcal{J}} s_m^{(j)}. \tag{5}$$

We report TaskComp (Task Completion) as the mean of per-instance scores over the evaluation set. The judge pool includes strong large language models (e.g., GPT, Gemini, DeepSeek, Grok).

Information Grounding. Judges are additionally shown both the ground-truth and predicted trajectories and emit a single overall groundedness scalar. For instance m, each judge $j \in \mathcal{J}$ outputs $g_m^{(j)} \in [0, 1]$ indicating the proportion of key ground-truth steps covered by the prediction. The final $\mathrm{InfoGrnd}_m$ is computed via the same trimmed-mean procedure as Task Completion (Equation 5; Prompt: Appendix B.2).

5 Experiment

5.1 Evaluation Results

As shown in Table 2, `Gemini 3.1 Pro` (Google DeepMind, 2026) attains the highest Average Score (0.615), followed by `Claude Opus 4.8` (0.596) and `GPT-5.5` (0.587). `Gemini 3.1 Pro` also leads Step Coherence (0.582), Order Consistency (0.453), and Merge Purity (0.659); `GPT-5.5` leads Recall (0.745) and Argument Similarity (0.667), while `Claude Opus 4.8` leads Precision (0.668) and Task Completion (0.682). The best Order Consistency remains below 0.5, indicating substantial headroom in cross-step tool use.

In contrast, compact/latency-oriented variants trade coverage and structural consistency for speed: for example, `Claude 4.5 Haiku` (Anthropic, 2025a) achieves high Precision (0.480) but much

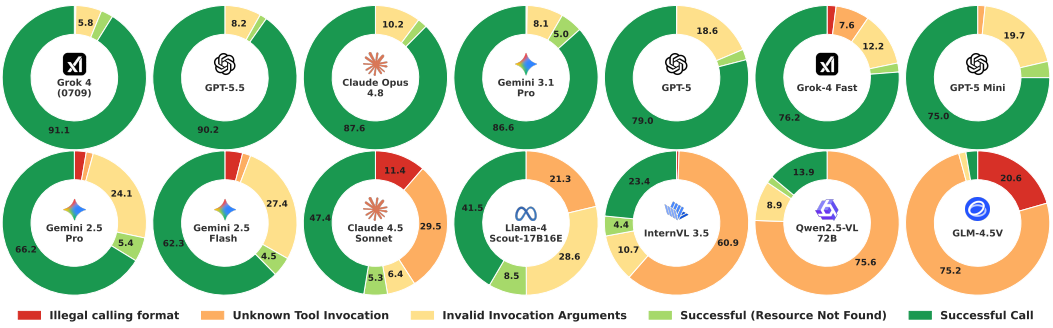


Figure 5: Per-call outcomes from replaying predicted trajectories through the MCP interpreter.

lower Recall, and its structure metrics remain modest, pointing to short, locally correct trajectories that under-utilize multi-threaded branches and often break multi-hop dependencies. Finally, `GLM 4.5v` (Hong et al., 2025), `Qwen2.5-VL-72B` (Bai et al., 2025b), and `InternVL 3.5` (Wang et al., 2025a) lag on both coverage and structure, suggesting that MLLMs SFT objectives favor pure vision–language understanding over instruction following and trajectory planning.

5.2 Per-Task Analysis

Figure 6 shows that model capabilities are far from uniform across selected tasks (details in Appendix B): curves cross frequently, indicating complementary strengths rather than a single, uniformly dominant system. `Grok-4` is weak on *Fruit Cal* (numerical calibration/unit handling) yet strong on *QR-code/Attraction* (optical character recognition (OCR) and points-of-interest (POI) chaining). `GPT-5` trails on *Vision Research* (fine-grained figure grounding) but dominates *Biology Cut* (precise localization/cropping). `Gemini 2.5 Pro` excels on *Travel Comparison* (web aggregation and itinerary synthesis). `Claude 4.5` lags on vision-dependent tasks (e.g., *Items Price*), suggesting comparatively weaker multimodal grounding; disparities align with differences in OCR robustness. In Appendix F, we present the performance of different models on each task, based on the evaluation metrics of M^3-Bench.

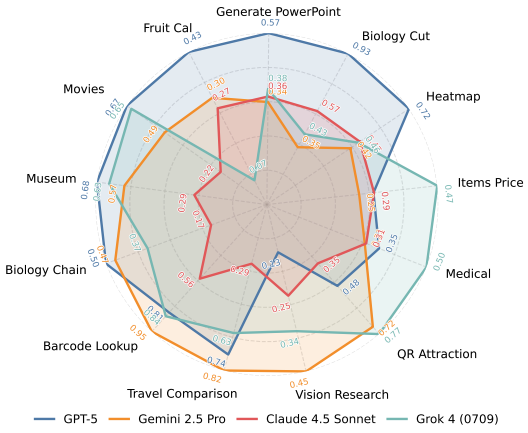


Figure 6: Average score of each selected task (details in Appendix B) in the representative models.

5.3 Analysis of Error Distributions

We assess per-call reliability by replaying each model's trajectory through the MCP interpreter and categorizing every `tool_call` into five exclusive outcomes (Figure 5): (i) *Illegal calling format*: the request cannot be parsed into a valid schema; (ii) *Unknown Tool Invocation*: the request targets a non-existent tool; (iii) *Invalid Invocation Arguments*: the tool exists but arguments are invalid (typically HTTP 400); (iv) *Successful (Resource Not Found)*: the call is valid yet the backend cannot locate the resource; and (v) *Successful call*: the tool executes and returns a normal result. These categories decouple schema conformance, tool selection, argument fidelity, and retrieval robustness.

Overall, higher-performing systems allocate a larger share to *Successful call* in Figure 5, with the GPT-5, Grok, and Gemini families exhibiting the strongest success rates. For `Qwen2.5-VL-72B` (Bai et al., 2025b) and `GLM-4.5V` (Hong et al., 2025), the shortfall appears driven less by multi-hop reasoning and more by tool-use hygiene: hallucination-triggered invocations of non-existent tools and weaker adherence to schema constraints (trajectory error analysis in Appendix Error: GLM-4.5v, Error: Qwen2.5-VL, Error: Internvl-3.5). This suggests that improvements in schema guidance and instruction following would yield larger gains than refining high-level reasoning alone. Although the Claude family is generally regarded as possessing strong capabilities in coding, mathematics,

and reasoning, its cross-modal prior knowledge is not particularly prominent (see Section E in Appendix). To further quantify this phenomenon, we introduce auxiliary vision-specific diagnostics in Appendix B.1 that measure whether planning failures are associated with missed perception triggers, weak visual grounding, or a vision bottleneck relative to non-visual planning steps.

6 Ablation Study

Ablation on human verification.

Our benchmark references are produced by the Executor–Judge loop with an optional *lightweight* human verification pass that audits scope and logical consistency. This pass is deliberately limited to (i) guiding the trajectory toward specialized MCP tools, (ii) removing sensitive artifacts (e.g., exposed API keys), and (iii) fixing minor multi-hop/multi-threaded inconsistencies. To quantify the effect of this optional step, we recompute all trajectory metrics using (a) the human-verified references (default setting in the main paper) versus (b) the raw MLLM-synthesized trajectories *before* the human pass as references. As shown in Table 3, the overall trends remain consistent: model rankings and relative gaps change only marginally, indicating that our evaluation is not overly sensitive to light human refinement, while the refinement improves rigor by filtering edge-case artifacts. *Note:* Table 3 and Figure 7 report only the six alignment-based metrics, excluding the two LLM-Judge metrics (TaskCp, InfoGr) held constant across these ablations.

Model	Recall	Prec	ArgSim	StepCoh	MergePur	OrdCons
GPT-5	.623|**.607**	.276|**.244**	.566|**.563**	.505|**.482**	.437|**.433**	.304|**.270**
Gemini 2.5 Pro	.492|**.481**	.463|**.429**	.312|**.413**	.454|**.444**	.395|**.404**	.238|**.229**

Table 3: **Human-verified** | **Raw** references. Bold numbers use raw MLLM-synthesized trajectories (before human verification) as references; non-bold numbers use the human-verified references reported in the paper.

Ablation on the sentence encoder. Our similarity-bucketed Hungarian alignment embeds each serialized tool call with a fixed sentence encoder E (Section 4.1) before cosine matching within tool buckets. We test four widely-used encoders: MPNet-base-v2, BGE-small-en-v1.5, BGE-base-en-v1.5, and E5-small-v2. Across the full evaluation set, the resulting scores and rankings are identical (even when rounded to 4 decimal places), suggesting that the alignment is robust to the specific choice of E under our canonical serialization and tool-bucketed matching.

Ablation on similarity thresholds τ_{weak} and τ_{strong}.

We analyze evaluation sensitivity to similarity thresholds in alignment. Recall that τ_{weak} gates feasible matching edges in the cost matrix (Eq. (1)), while τ_{strong} controls which matched pairs contribute to Argument Similarity (Eq. (2)). We sweep τ_{weak} and report average scores for representative models in Figure 7. When τ_{weak} is too high, the matching becomes overly strict and scores exhibit higher variance due to reduced feasible correspondences; when τ_{weak} is too low, the matching becomes overly permissive and can admit noisy argument pairs, slightly diluting alignment-based signals. Importantly, in the intermediate range $\tau_{\text{weak}} \in [0.4, 0.8]$, rankings remain stable and trends are consistent, indicating a robust operating regime; we use $(\tau_{\text{strong}}, \tau_{\text{weak}}) = (0.8, 0.6)$ as our default setting.

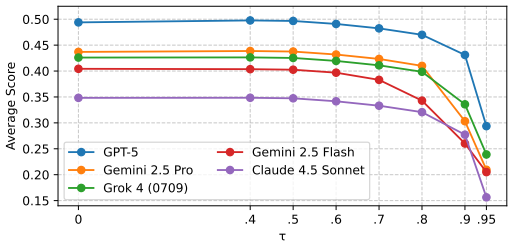


Figure 7: Sensitivity to the similarity threshold τ_{weak}. Scores and rankings of representative models remain stable for $\tau_{\text{weak}} \in [0.4, 0.8]$.

7 Conclusion

We introduced M^3-Bench for multimodal, multi-hop, and multi-threaded MCP trajectories, with similarity-bucketed alignment and structure-aware metrics. Across 14 MLLMs, higher Recall often coincides with lower Merge Purity and Order Consistency, while perception-dependent tasks show lower Argument Similarity even for top systems.

AI USE STATEMENT

LLMs generated benchmark trajectories and served as judges and evaluated systems, as described in the paper; the trajectories were verified by the authors. Generative AI also assisted with manuscript readability and code edits. The authors reviewed this work, tested the code, and take responsibility for the final paper and artifacts.

REFERENCES

Anthropic. Introducing the model context protocol. https://www.anthropic.com/news/model-context-protocol, 2024. Accessed: 2025-10-30.

Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, May 2025. Product announcement; system card available; no peer-reviewed venue acceptance announced as of 2025-11-10.

Anthropic. Claude haiku 4.5 system card. https://www.anthropic.com/claude-haiku-4-5-system-card, 2025a. Oct. 2025, official system card; accessed 2025-11-10.

Anthropic. Claude sonnet 4.5 system card. https://www.anthropic.com/claude-sonnet-4-5-system-card, 2025b. Oct. 2025, official system card; accessed 2025-11-10.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923. Covers Qwen2.5-VL-72B flagship variant.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2507.06261. Status: arXiv preprint; no peer-reviewed venue acceptance found as of 2025-11-10.

Zihan Dong, Zhixian Zhang, Yang Zhou, Can Jin, Ruijia Wu, and Linjun Zhang. Evaluating llms when they do not know the answer: Statistical evaluation of mathematical reasoning via comparative signals. *arXiv preprint arXiv:2602.03061*, 2026.

Yu Du, Fangyun Wei, and Hongyang Zhang. Anytool: Self-reflective, hierarchical agents for large-scale API calls. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11812–11829. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24h.html.

Shiqing Fan, Xichen Ding, Liang Zhang, and Linjian Mo. MCPToolBench++: A large scale ai agent MCP tool use benchmark. *arXiv preprint arXiv:2508.07575*, 2025.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, et al. Are: Scaling up agent environments and evaluations. *arXiv preprint arXiv:2509.17158*, 2025.

Xuanqi Gao, Siyi Xie, Juan Zhai, Shiqing Ma, and Chao Shen. MCP-RADAR: A multi-dimensional benchmark for evaluating tool-use capabilities in large language models. *arXiv preprint arXiv:2505.16700*, 2025. URL https://arxiv.org/abs/2505.16700.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Announced Feb. 19, 2026; accessed 2026-05-01.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. In *arXiv preprint arXiv:2412.17032*, 2024.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Z. Gong, and Lichao Sun. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Can Jin, Hongwu Peng, Qixin Zhang, Yang Zhou, Yujin Tang, Tong Che, and Dimitris N. Metaxas. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*, 2025a. doi: 10.48550/arXiv.2504.09772. URL https://arxiv.org/abs/2504.09772.

Can Jin, Hongwu Peng, Shiyu Zhao, Zhenting Wang, Wujiang Xu, Ligong Han, Jiahui Zhao, Kai Zhong, Sanguthevar Rajasekaran, and Dimitris N. Metaxas. Apeer: Automatic prompt engineering enhances large language model reranking. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW Companion '25, pp. 2494–2502, Sydney, Australia, May 2025b. Association for Computing Machinery. doi: 10.1145/3701716.3717574. URL https://doi.org/10.1145/3701716.3717574.

Can Jin, Yang Zhou, Qixin Zhang, Hongwu Peng, Di Zhang, Zihan Dong, Marco Pavone, Ligong Han, Zhang-Wei Hong, Tong Che, and Dimitris N. Metaxas. AIRL-S: Unifying reinforcement learning and search-based test-time scaling via adversarial inverse reinforcement learning. *arXiv preprint arXiv:2508.14313*, 2025c. doi: 10.48550/arXiv.2508.14313. URL https://arxiv.org/abs/2508.14313.

Thibault Le Sellier De Chezelles et al. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.pdf.

Xiao Liu et al. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Zhiwei Liu, Jielin Qiu, Shiyu Wang, Jianguo Zhang, Zuxin Liu, Roshan Ram, Haolin Chen, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. MCPEval: Automatic MCP-based deep evaluation for AI agent models. *arXiv preprint arXiv:2507.12806*, 2025. URL https://arxiv.org/abs/2507.12806.

Ziyang Luo, Zhiqi Shen, Wenzhuo Yang, Zirui Zhao, Prathyusha Jwalapuram, Amrita Saha, Doyen Sahoo, Silvio Savarese, Caiming Xiong, and Junnan Li. MCP-Universe: Benchmarking large language models with real-world model context protocol servers, 2025. URL https://arxiv.org/abs/2508.14704.

Meta AI. Llama 4 scout 17b-16e model card. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E, 2025. HF official model card; release Apr. 5, 2025; accessed 2025-11-10.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.12983.

Guozhao Mo, Wenliang Zhong, Jiawei Chen, Qianhao Yuan, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. Livemcpbench: Can agents navigate an ocean of mcp tools? *arXiv preprint arXiv:2508.01780*, 2025.

Model Context Protocol Working Group. Model context protocol (mcp) specification. https://modelcontextprotocol.io/specification/2025-06-18, 2025. Accessed 2025-10-22.

OpenAI. GPT-5 mini. https://platform.openai.com/docs/models/gpt-5-mini, 2025a. API docs; accessed 2025-11-10.

OpenAI. GPT-5 nano. https://platform.openai.com/docs/models/gpt-5-nano, 2025b. API docs; accessed 2025-11-10.

OpenAI. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025c. Aug. 2025, official system card; accessed 2025-11-10.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4020. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract-Conference.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yujia Qin et al. Toolllm: Facilitating large language models to master 16,000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410/.

Salesforce AI Research. MCP-Universe: Benchmarking large language models with real-world MCP servers. https://github.com/SalesforceAIResearch/MCP-Universe, 2025a. Accessed 2025-10-22.

Salesforce AI Research. MCP-Universe: A comprehensive framework for ai agent development and benchmarking. https://mcp-universe.github.io/, 2025b. Accessed 2025-10-22.

Scale AI. MCP-Atlas leaderboard/dataset overview. https://scale.com/leaderboard/mcp_atlas, 2025. Accessed 2025-10-22.

Scale AI Research. MCP-Atlas raises the bar for agentic evaluation. https://scale.com/blog/mcp-atlas, 2025. Accessed 2025-10-22.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2302.04761.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with ChatGPT and its friends in huggingface. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17580.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

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*, 2025b.

xAI. Grok 4 model card. https://x.ai/docs/model-cards/grok-4, 2025a. Initial public release July 9, 2025 ("0709"); accessed 2025-11-10.

xAI. Grok 4 fast model card. https://x.ai/docs/model-cards/grok-4-fast, 2025b. Official model card; accessed 2025-11-10.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiecao Chen. Toolhop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. URL https://arxiv.org/abs/2408.00765.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

Appendix Content

A Structure-Sensitive Metrics Overview 2
B Task Descriptions 2
B.1 Auxiliary Vision-Specific Diagnostics 6
B.2 LLM Prompts Used in Executor and Judge 8
B.3 Bucketed Hungarian Sample Pairs Based on Similarity 11
B.4 Human Study 11
C Token Budgets and Interaction Depth 12
D Trajectory Case Study 12
E Error Trajectory Distribution Analysis 17
F Per-tasks M^3-Bench Evaluation 21
G Ranking Stability 36

A Structure-Sensitive Metrics Overview

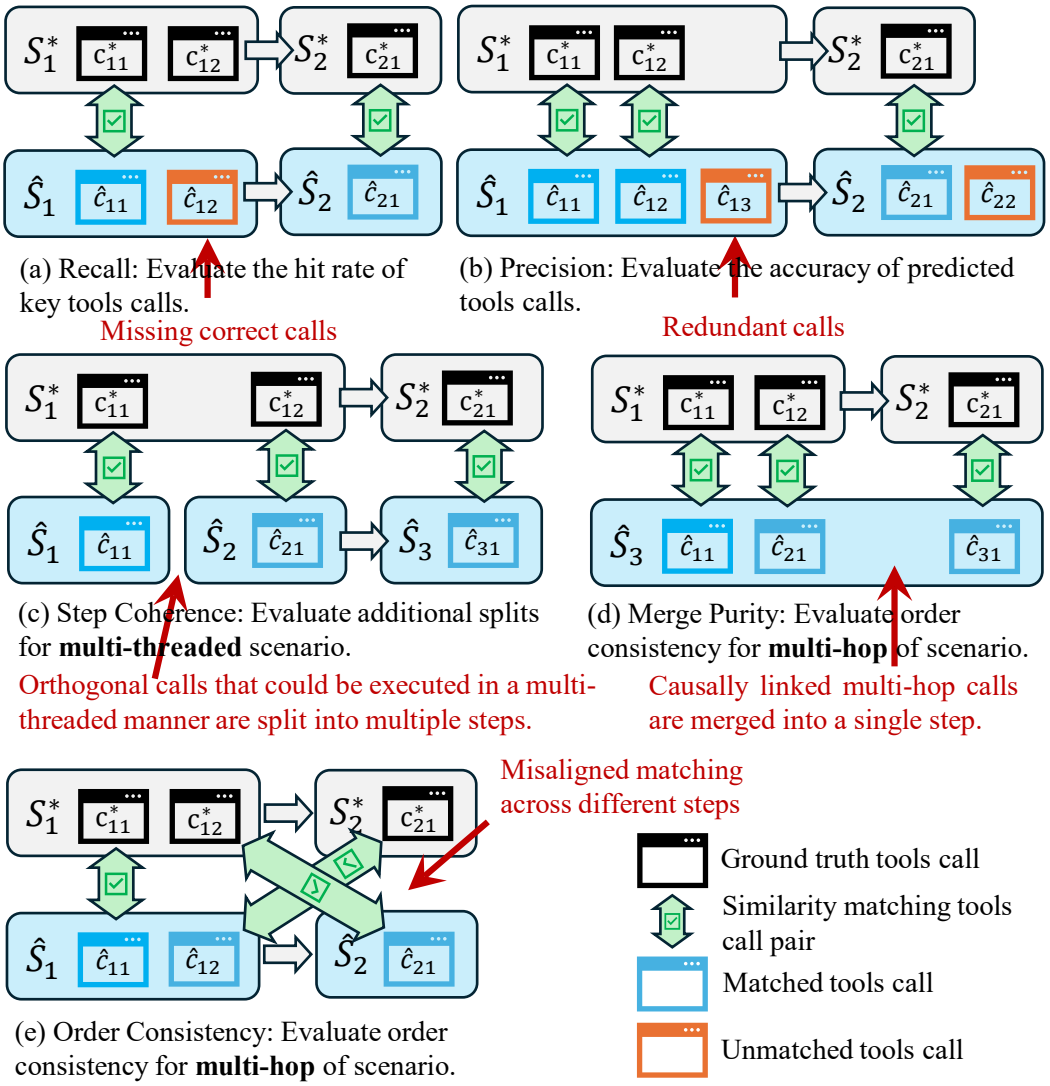


Figure 8: Our structure-sensitive metrics. where $S^\star_\ell$ and $\widehat{S}_{\hat{\ell}}$ denote the sets of calls in the ground-truth step ℓ and the predicted step $\hat{\ell}$, respectively, and $c^\star_{\ell,n^\star_\ell}$ and $\hat{c}_{\hat{\ell},n_{\hat{\ell}}}$ denote the $n^\star_\ell$-th ground-truth call in step ℓ and the $n_{\hat{\ell}}$-th predicted call in step $\hat{\ell}$.

B Task Descriptions

(0) Items Price Estimate prices for all visible retail items.

1. Identify product names from the image (OCR/visual grounding as needed).
2. Use Amazon MCP (search_products) to retrieve current prices.
3. Aggregate results into a concise table.

(1) Attraction Resolve the attraction, surface context, and plan around-visit options.

1. Use Wikipedia MCP to obtain the city/background of the attraction.

2. Call weather/get_weather for current conditions.
3. Use google-maps/places_nearby_search to find nearby restaurants.

(2) Movies Identify the movie and retrieve showtimes contextually.

1. Use OCR to read the movie title in the image.
2. Query TMDB to disambiguate the title.
3. Search cinemas near New York showing the resolved film via Google Maps MCP.

(3) Vision Research Bridge detected fauna to literature snippets.

1. Detect animals and identify two distinct categories from detections.
2. For each, call paper_search/search_arxiv with a targeted query.
3. Download and read the paper; extract the shortest Abstract sentence containing the animal name.

(4) Camping Plan a 3-day national-park camping itinerary with verified amenities and logistics.

1. Use nationalparks/getCampgrounds and nationalparks/getAlerts to select an open campground with ≥ 3 amenities and no active alerts.
2. Call nationalparks/getVisitorCenters and pick the closest visitor center; record hours.
3. Use google-maps/compute_route, distance_matrix, and elevation to compute routes and elevation; compile the itinerary.

(5) Fish Plan a New Jersey fishing trip for the detected species.

1. Use wiki/summary to summarize the detected fish species.
2. Use wiki/search to search for the top spot for fishing in NJ recently.
3. Call weather/get_weather for local conditions at the spot.
4. Use google-maps/compute_route to plan a route from a default start.
5. Use google-maps/places_nearby_search to search for restaurants and bait shops.

(6) Drug Compare medicines for a given condition safely and conveniently.

1. Look up drug facts via fda_drug_lookup MCP.
2. Search similar products via Amazon (search_products).
3. Create a comparison list using Excel MCP.

(7) Hazard Document construction-safety hazards with visual annotations and slides.

1. Run a detector to find hardhats, vests, ladders, and unguarded edges.
2. Use imagesorcery-mcp/draw_circles to annotate detections.
3. Create a presentation with ppt/create_presentation, ppt/add_slide and ppt/manage_image.

(8) Food Reddit Wiki Excel Quantify food popularity and descriptions, then export to Excel.

1. For each food, query Reddit MCP for the hottest post and extract comment count.
2. Use Wiki MCP to get a brief description.
3. Create a workbook and write two columns (Food, Wikipedia Description) via Excel MCP; save the file.

(9) MedExcel Summarize PubMed signal into key terms in Excel.

1. Use pubmed_search to retrieve the first three titles for the drug/disease.
2. Double-check whether each title is semantically relevant to the queried drug/disease.
3. Write the drug/disease name and the relevant paper count to an Excel file under ./media and save.

(10) Barcode Lookup Identify the book via barcode and fetch bibliographic data.

1. Decode the barcode using pyzbar.
2. Query openlibrary_mcp to return the book record.

(11) Biology Cut Produce one image per detected animal instance.

1. Detect all animals in the image.
2. Crop each detection to an individual image.

(12) Fruit Cal Assess daily calories and potential long-term risks from fruit-only meals.

1. Use Nutrition MCP to estimate total daily calories from the shown fruits.
2. Use Healthcare MCP to assess long-term health concerns of such a diet.

(13) Medical Recommend safe medicines with purchase links conditioned on ICD-10.

1. Use lookupICDCode MCP to validate the ICD-10 code.
2. Use fda_drug_lookup and clinical_trials_search to check safety/compatibility.
3. Suggest appropriate options via Amazon MCP with links.

(14) QR Attraction Infer location from a QR-sourced page and report air quality and guidance.

1. Use a QR scanner tool to extract the URL; fetch the primary image.
2. Infer the location and geocode it via Google Maps MCP.
3. Query google-air/current_conditions for AQI and health recommendations.

(15) Routine Plan a multi-stop route with a supermarket stop and choose the fastest option.

1. Geocode the start, the image landmark, and the destination via Google Maps MCP.
2. Search for nearby supermarkets as waypoints.
3. Compare ≥ 3 candidate routes (compute_route or directions) and select the fastest.

(16) Biology Chain Evaluate whether detected organisms can form a plausible food chain.

1. Detect all organisms; blur any human instances.
2. Use Wikipedia to retrieve ecological roles (diet, trophic level).
3. Check whether roles form a valid chain.

(17) Generate PowerPoint Recreate a title slide with style and optional Q&A enhancement.

1. OCR the slide and segment into blocks; estimate font color/size/style.
2. Call create_presentation to get presentation_id; use create_slide_from_template to build the slide.
3. If a question is present, fetch and summarize Reddit opinions for a Q&A slide.
4. save_presentation to output PPTX.

(18) Travel Comparison Choose the cleaner destination between two landmarks and suggest nearby venues.

1. Recognize both landmarks and geocode each via Google Maps MCP.
2. Use google-air/current_conditions to compare AQI and pollutant levels.
3. Recommend the cleaner destination; list nearby cinemas and hotels.

(19) Heatmap Generate an air-pollution heatmap visualization for the depicted city.

1. Geocode the city via Google Maps MCP.
2. Render a static base map (static_map) with zoom=10.
3. Overlay PM2.5 heatmap tiles via air-quality/heatmap_tile (indexType = US_AQI).

(20) Museum Retrieve a Met object and attach a concise Wikipedia summary.

1. Call metmuseum-mcp/list-departments and select the best-aligned departmentId.
2. Call metmuseum-mcp/search-museum-objects (hasImages=true, departmentId) and select an objectId.
3. Call metmuseum-mcp/get-museum-object to get title, classification, and image URL.
4. Use wiki/search then wiki/summary to obtain a short cultural summary.

(21) Research Station Relate solar flare events to nearby research stations for the inferred location.

1. Use nasa-mcp/get_notifications to filter M5+ flares within the specified window.
2. For each event date, call get_solar_flare and get_coronal_mass_ejection; store flux/-peak_time and CME IDs/types.
3. Infer photo location; use google-maps-mcp/places_text_search and place_details to list top research stations and compile a table.

(22) Hugeicons Map salient objects to Wikipedia titles and validated icons.

1. Use dinox-mcp/detect-all-objects and keep detections with area $>$ 10,000.
2. For each object, call wiki/search_wikipedia to get an article title.
3. Call hugeicons/search_icons and hugeicons/list_icons; cross-validate up to two icon names per object.
4. Output a JSON table with Object, Wikipedia Title, and Icon Names.

(23) Crypto Purchase Estimation Estimate how many main products 100 units of the most volatile currency can buy.

1. Use dinox-mcp/detect-all-objects and select the object with the largest bounding box as the main product.
2. Query amazon/search_products for price/title; compute 24h volatilities from okx-mcp/get_candlesticks across candidate instruments and pick the most volatile.
3. Get current price and 24h volume via okx-mcp/get_price.
4. Compute purchasing power and estimated items purchasable; return structured JSON.

(24) Stock Growth Estimation Estimate the short-term growth potential of the companies and annotate the image with the final result.

1. Use yahoo-finance/get_stock_info to retrieve real-time prices for a group of candidate company tickers and keep the two highest-priced companies for further analysis.

2. For selected companies, call yahoo-finance/get_historical_stock_prices over 2024-10-01 to 2024-10-03, compute the average of their daily Close prices, and keep the company with higher average for further analysis.
3. Call yahoo-finance/get_recommendations for the final selected company and compute the upgrade ratio. Then compute the short-term growth percentage using the formula: Growth Percentage = $\frac{\text{Real-time Price} - \text{Average Close Price}}{\text{Average Close Price}} \times 100 + (\text{Upgrade Ratio} \times 10)$.
4. Use imagesorcery-mcp/draw_texts to write the computed growth estimate onto the input image, and return the resulting ticker, its price metrics, the upgrade ratio, the estimated growth, and the annotated image.

(25) FIPE Car Variant Lookup Identify the vehicle, determine its 20xx FIPE variant, and summarize its fuel type.

1. Identify the vehicle and extract both the brand name and model keywords. Query car-price/get_car_brands for the complete list of supported brands. Call okx/get_exchange_rate to obtain the current BRL→USD and BRL→BTC conversion rates.
2. If the brand exists in the supported list, then use its brand name and model keywords, call car-price/search_brand_model_price to retrieve all matching FIPE entries, filter these results to keep only the 20xx variant.
3. Use wiki/summary to look up the fuel type of the 20xx variant, and provide a short description based on the search result.

(26) Class Demand Estimation Estimate class demand, compare it with the detected object count, and determine whether the class budget is sufficient.

1. Use Math-MCP tools math/mean and math/median to compute median and mode from the collected survey statistic {6, 8, ...}. Use dinox-mcp/detect-all-objects to detect all objects and identify the most frequently appearing object category. Compare the survey median, the survey mode, and the detected top object count, then take the largest among them as the estimated demand.
2. Use amazon/search_products to query the detected top object category, obtain the first returned product, and compute the total cost as the ceiling of the estimated demand multiplied by the item price.
3. If the total cost is within the class budget of xxx dollars budget, call Reddit-MCP-Server/search_hot_posts to fetch the first recent hot post and wiki/summary to obtain a short description for the detected object category.

(27) NixOS Diagnostic Perform a staged NixOS diagnostic workflow using OCR, NixOS tools, and Wiki tools.

1. Use ocr/perform_ocr exactly once to extract the package names appearing in the code on the image. These package names will be used in the following steps.
2. Take the first five package names in their original order and call nixos/ nixos_search once for each of them to check whether they exist in NixOS. If fewer than three of these packages exist, stop the workflow and report.
3. Otherwise, collect the first three package names that do exist. Call nixos/ nixos_info on services.openssh.enable. Then call nixos/home_manager_search with the query "zsh" and keep the first three returned options; choose the first one and call nixos/home_manager_info. Finally, call wiki/summary once for each of the three existing packages to provide brief background information.

B.1 Auxiliary Vision-Specific Diagnostics

The error trajectory analysis suggests that a subset of failures is not primarily caused by long-horizon planning alone, but by earlier visual understanding errors that propagate into downstream tool selection and argument construction (e.g., incorrect landmark recognition leading to incorrect

geolocation and follow-up retrieval calls). To make this phenomenon measurable, we add three auxiliary diagnostics on top of the same similarity-bucketed Hungarian alignment used in Section 4.1.

Scope and implementation convention. We reuse the per-trajectory matched pair set M and matched similarities S_{ij} produced by the tool-bucketed Hungarian alignment with weak threshold τ_{weak} (Section 4.1). In the implementation used for this appendix, the auxiliary metrics are computed with $\tau_{\text{weak}} = 0.6$, and are aggregated only over tasks where the corresponding signal is applicable (e.g., only tasks containing perception calls for Perception Recall). The script also defines a fixed tool taxonomy for *perception* and *grounding* tools (including explicit tool names and prefixes), and then treats their union as *vision* calls.

For a ground-truth call index i, define the matched similarity score

$$s_i = \begin{cases} S_{ij}, & \text{if } (i,j) \in M, \\ 0, & \text{otherwise,} \end{cases}$$

i.e., unmatched calls contribute zero.

Perception Recall. Perception Recall measures whether a model correctly triggers required *perception* tools (e.g., OCR, detection, image processing) when such calls are present in the ground-truth trajectory. Low values indicate that the model may fail to "see" and therefore never enters the correct visual-processing branch.

Let $\mathcal{I}_{\text{perc}}$ be the set of ground-truth call indices whose tools are labeled as perception tools. We define

$$\text{PerceptionRecall} = \frac{|\{i \in \mathcal{I}_{\text{perc}} \ : \ \exists j, \ (i,j) \in M\}|}{|\mathcal{I}_{\text{perc}}|},$$

with the convention $\text{PerceptionRecall} = 1$ when $|\mathcal{I}_{\text{perc}}| = 0$ for a task (not applicable). The reported score is the average over tasks that contain at least one perception call.

Grounding Similarity. Grounding Similarity measures the quality of *grounding* calls (e.g., search/location/product grounding calls) by averaging matched similarity over all required grounding calls, with missed calls counted as zero. Compared with the global ArgSim in Sec. 4.3, this diagnostic is restricted to grounding calls and is recall-aware by construction.

Let $\mathcal{I}_{\text{grd}}$ be the set of ground-truth grounding-call indices. We define

$$\text{GroundingSimilarity} = \frac{\sum_{i \in \mathcal{I}_{\text{grd}}} s_i}{|\mathcal{I}_{\text{grd}}|},$$

with the convention $\text{GroundingSimilarity} = 1$ when $|\mathcal{I}_{\text{grd}}| = 0$ for a task (not applicable). The reported score is the average over tasks that contain at least one grounding call.

Visual Reasoning Ability. Visual Reasoning Ability is a task-level binary diagnostic that tests whether a model's average performance on *vision* calls (perception + grounding) is at least as strong as its average performance on non-visual planning calls within the same task.

Let

$$\mathcal{I}_{\text{vis}} = \mathcal{I}_{\text{perc}} \cup \mathcal{I}_{\text{grd}}, \qquad \mathcal{I}_{\text{plan}} = \{1, \ldots, N_{\text{gt}}\} \setminus \mathcal{I}_{\text{vis}},$$

where N_{gt} is the number of ground-truth calls in the task. Define the average matched-similarity scores

$$\bar{s}_{\text{vis}} = \frac{1}{|\mathcal{I}_{\text{vis}}|} \sum_{i \in \mathcal{I}_{\text{vis}}} s_i, \qquad \bar{s}_{\text{plan}} = \frac{1}{|\mathcal{I}_{\text{plan}}|} \sum_{i \in \mathcal{I}_{\text{plan}}} s_i.$$

We then define the task-level indicator

$$\text{VRA} = \mathbb{I}[\bar{s}_{\text{vis}} \geq \bar{s}_{\text{plan}} \ \wedge \ \bar{s}_{\text{vis}} > 0.4] \, .$$

The additional $\bar{s}_{\text{vis}} > 0.4$ condition avoids trivial passes in near-zero regimes (e.g., both visual and planning performance are uniformly poor). If a task has no planning calls, the metric is treated as N/A for that task and excluded from the aggregate. The reported score is the average VRA over applicable tasks.

Results. Table 4 reports the three auxiliary diagnostics for the evaluated models ($\tau_{\text{weak}} = 0.6$). The results support the qualitative error analysis: some models fail mainly because they do not reliably trigger the required perception/grounding branch, while others exhibit stronger perception coverage but weaker grounding fidelity. In particular, GPT-5 has the highest Perception Recall (0.826), Gemini-2.5 Pro has the highest Grounding Similarity (0.582), and Grok-4 (0709) has the highest Visual Reasoning Ability (0.455) among the tested models.

Table 4: Auxiliary vision-specific diagnostics on M3-Bench CV-related evaluation. **Perception Recall** measures whether required perception tools are correctly triggered. **Grounding Similarity** evaluates the similarity of grounding-call matches (with missed required grounding calls contributing zero). **Visual Reasoning Ability** reports the fraction of applicable tasks where average performance on vision calls is at least as strong as on non-visual planning calls (with an additional $\bar{s}_{\text{vis}} > 0.4$ gate). **Avg Score** is the arithmetic mean of the three diagnostics. Higher is better for all metrics.

Model	Auxiliary Vision Diagnostics			Avg Score
	Perception Recall	Grounding Similarity	Visual Reasoning Ability	
GLM-4.5v	0.016	0.008	0.000	0.008
Qwen2.5-VL-72B	0.328	0.159	0.276	0.254
Claude-4.5 Haiku	0.360	0.210	0.256	0.275
InternVL-3.5	0.410	0.213	0.284	0.302
Gemini-2.5 Flash Lite	0.362	0.355	0.321	0.346
Llama-4-Scout-17B-16E	0.471	0.357	0.343	0.390
GPT-5 Nano	0.537	0.409	0.315	0.420
Claude-4.5 Sonnet	0.577	0.312	0.407	0.432
Grok-4 Fast	0.541	0.334	0.435	0.437
Gemini-2.5 Flash	0.782	0.374	0.433	0.530
GPT-5 Mini	0.653	0.532	0.414	0.533
Gemini-2.5 Pro	0.672	0.582	0.388	0.547
Grok-4 (0709)	0.781	0.521	0.455	0.586
GPT-5	0.826	0.544	0.448	0.606

B.2 LLM Prompts Used in Executor and Judge

Prepare (Tool Planning) Following the planning protocol in Section 3.1, the Executor first selects (but does not execute) the tool set for the current step. The exact prompt template used for this stage is given below.

Prepare (Tool Planning)

```
   You are an assistant with MCP tool invocation capability. You must
call one or more MCP tools, possibly multi-hop, to complete the task.
Consider necessary steps as much as possible, consider completed steps
 in conjunction with conversation history, and skip unnecessary steps.
 Extra steps may reduce your final score
   This is step {i} of {self.max_step}.
   Now decide which tools to use for this step.
   Output EXACTLY in the following format (no extra text):

     "server_name/tool_name: <description>
      Args: <arguments (type): description>
      Returns: <returns(type): description>"

   <plan for how to use>

   Here are the MCP tools for you. All tools list (name: description):
   [the full tool catalog is appended programmatically]
```

Process (Controlled Execution) Given the selected tools, execution proceeds under controlled, JSON-only invocation with bounded parallelism as specified in Experiment Section 3.1. The prompt below instantiates that protocol.

Process (Controlled Execution)

```
   Now execute the selected tools for this step. Return STRICT JSON
ONLY with a single top-level key 'tool_calls' containing 1 to {self.
max_concurrent} items. Consider necessary steps as much as possible,
consider completed steps in conjunction with conversation history, and
 skip unnecessary steps. Extra tool calls may reduce your final score
   Each item must be an object: {"name": "server_name/tool_name", "
arguments": { ... }}.
   Do NOT include markdown fences or any extra text or special
delimiters and tokens.
   If a tool requires images/files, include the appropriate arguments
per the tool's schema. You may use the image pool paths listed below
if applicable.
   Selected tools (name: description):
   [The selected tools list is appended programmatically]
   [if available, image pool paths are appended]
```

Stop/Continue A binary stopping criterion gates the transition to summarization, as defined in Section 3.1. The following prompt enforces a strict yes/no decision.

Stop/Continue

```
   If the task can be considered complete based on current tool
results and question, here is the original question: {last_user},
   answer 'yes'. Otherwise 'no'.
   Answer strictly with 'yes' or 'no', no extra words.
```

Final Reply When the stopping criterion is met, the agent emits only the final natural-language answer (Section 3.1). The corresponding prompt template is provided below.

Final Reply

```
   You are finalizing the conversation. Produce ONLY the final answer
in natural language.
   Do NOT include any <tool_call> tags or mention tools explicitly. Be
 concise and accurate,
   relying on the prior tool results contained in the conversation.
```

Judge The Judge independently audits causal order and factual correctness, per Experiment Section 3.1. The prompt below operationalizes this decision.

Judge

```
   You are a strict task finisher. Read the provided interaction
history and decide if the task is completed.
   Very important! If there is an incorrect causal order in multi-step
 MCP tools calls, or if the information returned by the MCP tools call
 is clearly wrong, but the call chain proceeds smoothly due to LLM
hallucinations, it should be judged as an error.
   Output EXACTLY one of: success or fail. No other words.
```

Summarize Upon success, the Judge reconstructs an optimal trajectory by merging/removing calls and grouping parallel operations (Experiment Section 3.1). We provide the standard summarization prompt below.

Summarize

```
   You will summarize the optimal trajectory of useful MCP tool calls
to solve the task.
   Rules:
   - Remove failed or unnecessary tool calls.
   - Group calls by step; within a step, multiple calls are in
parallel (array order arbitrary).
   - Operations that can invoke the MCP tools call in parallel were
mistakenly executed across multi-hop --- they should be merged into a
single step.
   - Preserve only name and arguments for each call.
   - Output STRICT JSON only. No markdown, no comments, no extra text.
   - If you cannot determine a valid trajectory, output exactly the
string fail (no JSON).
```

Task Completion For end-to-end task completion, we use a four-judge trimmed-mean score defined in Experiment Section 4.4. The prompt below elicits the required scalar.

Task Completion

```
   You are a strict and professional academic reviewer tasked with
evaluating task completion when a model uses MCP tools.
   Judge solely based on the visible prediction content; do not use or
 allude to any ground truth.
   Scoring rule: output a single scalar in the form \\boxed{S}, and
nothing else. S is a float in [0,1].
   Rules:
   - 1) Planning (0-0.3): Clearly presents the MCP call process and
plan; complete and organized.
   - 2) Process (0-0.3): MCP calls obtain desirable tool feedback and
intermediate results; tool use is effective.
   - Final result (0-0.4): Degree of task completion (partial credit
allowed); final output addresses the task.
   - Please provide an overall score from 0 to 1 combining the three
parts, and output only \\boxed{S}.
```

Information Grounding Step-level groundedness is evaluated against the reference steps as defined in Experiment Section 4.4. The prompt below collects the coverage-only score.

Information Grounding

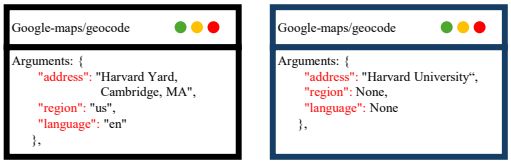
```
   You are a strict and professional academic reviewer responsible for
 evaluating information grounding.
   Task: compare the reference steps (ground truth steps) with the
predicted steps and determine whether the prediction includes all key
steps from the reference.
   Rules:
   - Coverage-only criterion: equivalence/paraphrase/minor order
changes count as covered;
   - Extra steps are not penalized;
   - Missing or clearly deviating key steps are penalized linearly by
the proportion of missing steps relative to the total reference steps;
   Output format: output only a scalar in the form \\boxed{G}, where G
 is a float in [0,1].
```

B.3 BUCKETED HUNGARIAN SAMPLE PAIRS BASED ON SIMILARITY

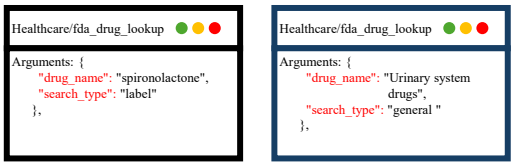
To make the alignment intuition concrete, Figure 9 shows three representative call pairs after canonicalization and sentence-encoder cosine scoring. Within each tool bucket, only pairs clearing a weak screen τ_{weak} are eligible for Hungarian assignment; those above a stronger bar τ_{strong} are treated as high-confidence matches. The high/medium/low examples illustrate how semantically close arguments receive credit despite surface variation, while dissimilar pairs are withheld—clarifying how our *Argument Similarity* feeds the bucketed matching used in Section 4.1 and, downstream, the structure-aware metrics.



(a) High similarity



(b) Medium similarity



(c) Low similarity

Figure 9: Illustrative call-level similarity under our sentence-encoder + cosine metric with bucketed Hungarian matching: three representative pairs exhibiting high, medium, and low similarity. This visualization contextualizes *Argument Similarity* and the weak/strong thresholding used for alignment (see Section 4.1; cf. related agent evaluation protocols.

B.4 HUMAN STUDY

We conducted a small-scale user study with 30 Ph.D. and researchers working with Model Context Protocol (MCP) tools to validate two design choices: (i) the use of human prompts and human judges (see Section 4.4) for obtaining an optimal trajectory; and (ii) the similarity thresholding used together with the bucketed Hungarian (Kuhn–Munkres) matching (see Section 4.1).

Preference for Optimal Trajectory. Each participant received 20 randomly sampled pairs of trajectories (the original MLLM-produced trajectory versus the human-optimized *Best Trajectory*), with order randomized and content blinded. Participants chose the better one per pair based on faithfulness and procedural soundness. Overall, the Best Trajectory was preferred in 98.16% of cases; under our pairwise setup this corresponds to 589/600 selections (98.17%).

Calibration of similarity thresholds. From the evaluation pool, we selected 40 ground-truth/predicted call pairs and asked participants to categorize them as *strong*, *medium*, or *weak* correlation (majority vote). We then swept the similarity thresholds and found that $(\tau_{\text{strong}}, \tau_{\text{weak}}) = (0.8, 0.6)$ produced the closest agreement with human labels.

C Token Budgets and Interaction Depth

Model	Input Tok.	Output Tok.
Qwen2.5-VL-72B	0.362	0.008
InternVL 3.5	0.069	0.002
Claude 4.5 Haiku	0.118	0.007
Claude 4.5 Sonnet	0.408	0.029
GPT-5 Nano	0.002	0.001
Llama-4-Scout-17B16E	0.345	0.008
Grok-4 Fast	0.210	0.028
Gemini 2.5 Flash Lite	0.477	0.010
Gemini 2.5 Flash	0.621	0.032
GPT-5 Mini	0.125	0.014
Grok 4 (0709)	0.132	0.027
Gemini 2.5 Pro	0.606	0.031
GPT-5	0.252	0.037

Table 5: Average token cost per task for MLLMs in millions.

Model	Avg. Steps	Avg. Calls
Qwen2.5-VL-72B	4.29	12.30
InternVL 3.5	3.65	9.57
Claude 4.5 Haiku	0.80	2.73
Claude 4.5 Sonnet	2.74	6.99
GPT-5 Nano	5.13	17.97
Llama-4-Scout-17B16E	4.72	13.26
Grok-4 Fast	2.91	5.93
Gemini 2.5 Flash Lite	2.73	7.06
Gemini 2.5 Flash	4.04	7.12
GPT-5 Mini	3.72	12.56
Grok 4 (0709)	3.21	7.64
Gemini 2.5 Pro	3.04	6.95
GPT-5	3.55	14.77

Table 6: Average interaction steps and tool calls per task.

Table 5 reports average per-task input and output token budgets, while Table 6 summarizes the average interaction depth per task. Larger models often pursue longer, multi-hop plans: `GPT-5` (OpenAI, 2025c) and `GPT-5 Nano` (OpenAI, 2025b) average (3.55 / 14.77) and (5.13 / 17.97) (steps / tool calls), indicating frequent retries and non-trivial orchestration; `Llama-4-Scout-17B16E` (Meta AI, 2025) and `Qwen2.5-VL-72B` (Bai et al., 2025b) show similarly heavy tool use. By contrast, `Gemini 2.5 Pro` (Comanici et al., 2025) (3.04 / 6.95), `Gemini 2.5 Flash` (Comanici et al., 2025) (4.04 / 7.12), and `Grok-4 Fast` (xAI, 2025b) (2.91 / 5.93) reach goals with fewer, more precise calls, consistent with their higher detection precision. Very short dialogues (e.g., `Claude 4.5 Haiku` (Anthropic, 2025a), `GLM 4.5v` (Hong et al., 2025)) reflect under-coverage rather than efficiency.

D Trajectory Case Study

Format. We present each trajectory using the two-column key–value layout defined below. Labels in the left column are typeset in monospace. `IMG` denotes the input image and `Q` the task query. `Step k` indicates the k-th multi-hop step in the execution. Within a step, dashed separators denote parallel (multi-threaded) tool calls. Each call reports `Tool` (invoked tool), `ARG` (arguments), and `OUT` (returned output). The final answer is summarized under `Final Reply`.

Object detection with per-instance cropping; parallel crop operations produce one image per detected animal. See (11) Biology Cut.

Example: Object detection and cropping

IMG	
Q	`Could you spot every creature in the picture and give me individual cutouts for each?`
Step 1	
Tool	`dinox-mcp/detect-all-objects`
ARG	`imageFileUri=./media/png; includeDescription=false`

Continued on next page

Example: Object detection and cropping (cont.)

OUT	Objects detected: horse (1), dog (2), elephant (1). Bboxes use {xmin, ymin, xmax, ymax}; coordinates describe exact positions and spatial relations.
Step 2	
Tool	imagesorcery-mcp/crop
ARG	input_path=./media/png; x1=473; y1=153; x2=1079; y2=952; output_path=./media/elephant.png
OUT	Image crop completed; saved as ./media/elephant.png.
Tool	imagesorcery-mcp/crop
ARG	input_path=./media/png; x1=1105; y1=364; x2=1428; y2=1015; output_path=./media/horse.png
OUT	Image crop completed; saved as ./media/horse.png.
Tool	imagesorcery-mcp/crop
ARG	input_path=./media/png; x1=61; y1=671; x2=359; y2=1024; output_path=./media/dog1.png
OUT	Image crop completed; saved as ./media/dog1.png.
Tool	imagesorcery-mcp/crop
ARG	input_path=./media/png; x1=338; y1=682; x2=531; y2=1024; output_path=./media/dog2.png
OUT	Image crop completed; saved as ./media/dog2.png.

Final Reply

Found 4 creatures: 1 elephant, 1 horse, 2 dogs. Each has a precise bbox ({xmin, ymin, xmax, ymax}, origin at top-left) and a cropped cutout under ./media/*.png.

Fish Species identification with geospatial planning: Wiki species summary, suitable local spot, weather, routing, and nearby POIs for an executable trip plan. See (5) Fish.

Example: Saltwater fish ID and NJ trip planning

IMG	
Q	From the NJ fish photo, can you ID the species with a brief overview, pick a recent good spot to try, check the current weather there, map a route from the usual starting point, and list nearby restaurants and a bait shop within about 5 km?
Step 1	
Tool	wiki/summary
ARG	title="American eel"
OUT	Species identified as American eel (*Anguilla rostrata*); catadromous, slender body, continuous dorsal/anal fin, nocturnal feeder common along Atlantic coast and NJ estuaries.
Tool	wiki/search
ARG	query="Sandy Hook fishing New Jersey"; n=5
OUT	Sandy Hook (Gateway NRA, near Highlands, NJ) highlighted as productive surf/shore area for eels and other inshore species.
Step 2	
Tool	weather/get_weather
ARG	location="Highlands, NJ"; units="us"
OUT	Current conditions around Highlands, NJ show mild temperatures (∼52.2°F) and light winds (∼4.1 mph), suitable for an evening shore session.
Step 3	
Tool	google-maps/compute_route
ARG	origin={latLng:{40.7357,-74.1724}}; destination={latLng:{40.4333,-73.9885}}; travelMode=DRIVE; fields=routes.distanceMeters,...
OUT	Route from Newark, NJ to Sandy Hook, NJ computed; distance ∼44 miles, duration ∼1 hour via Garden State Parkway and NJ-36 S.
Step 4	
Tool	google-maps/places_nearby_search
ARG	location={40.4333,-73.9885}; radiusMeters=5000; includedTypes=["restaurant","fishing_store"]; languageCode="en"; regionCode="US"
OUT	Nearby options include several waterfront restaurants (e.g., Bahrs Landing, The Proving Ground, Off The Hook, One Willow) and bait/tackle shops within 5 km.

Final Reply

Species: American eel (*Anguilla rostrata*); catadromous, active at night in tidal creeks/marshes. Spot: Sandy Hook, NJ (Gateway NRA near Highlands). Weather: calm, cool, fishable. Route: Newark→Sandy Hook = 44 mi / ∼1 h. Amenities: multiple waterfront restaurants and local bait & tackle within ∼5 km. Plan: fish evening high/outgoing tides along bayside creek mouths and rock structures; use bloodworms or cut bait; register with NJ Saltwater Recreational Registry and follow current regulations.

Example: Recreating Slide through Screenshot

IMG



Q From the screenshot of the first slide, pull out all visible text grouped into blocks (title, subtitle, footer if any) with rough font size/color/style guesses, recreate it as a new presentation with a matching title page, and if any line is a question, add a short Answer made from a few relevant community takes, then save it.

Step 1

Tool ocr/perform_ocr
ARG input_data=./media/png; language="eng"; config="-oem 3 -psm 6"
OUT Extracted multi-line title and subtitle text; detected embedded question about Reddit cartoons.

Tool Reddit-MCP-Server/search_hot_posts
ARG subreddit="cartoons"; limit=3
OUT Fetched top hot threads: "fav purple character?", "Ruh Roh Raggy", "Choo choo I guess...".

Step 2

Tool ppt/create_presentation
ARG id="pres"
OUT Initialized new PowerPoint deck with modern blue theme.

Step 3

Tool ppt/create_slide_from_template
ARG template_id="title_slide"; presentation_id="pres"; content_mapping={title, subtitle, question, answer}
OUT Title + subtitle mapped; question block appended for Q&A enhancement.

Tool ppt/manage_text
ARG operation="add"; slide_index=0; text="Question: what are latest hot thread in reddit/cartoon?"; font_size=20; bold=True; color=(0,120,215)
OUT Inserted question line below subtitle with bright blue accent.

Tool ppt/manage_text
ARG operation="add"; slide_index=0; text="- fav purple character?..."; font_size=20; bold=True; color=(0,102,204)
OUT Added Reddit answer list under question; auto-fit layout maintained.

Step 4

Tool ppt/save_presentation
ARG file_path=./media/Recreated_Slide.pptx; presentation_id="pres"
OUT File saved successfully; final slide index 0 complete.

Final Reply

```
Recreated title slide for a presentation on multi-modal vision-language models.
Title (large, bold, blue #0078D7); Subtitle (medium, italic, gray #444444).
Detected question: "what are latest hot thread in reddit/cartoon?".
Reddit answers inserted as Q&A section:
1) fav purple character?... 2) Ruh Roh Raggy... 3) Choo choo I guess...
Presentation ID: pres. Final file: Recreated_Slide.pptx.
Slide successfully rebuilt with question and answers in matching blue theme.
```

Barcodelookup Barcode decoding followed by OpenLibrary lookup for bibliographic metadata. See (10) Barcode Lookup.

Example: Book Identification via Barcode Lookup

IMG



Q From the back-cover barcode, can you identify the book and provide its title, author, publisher, year, and page count?

Step 1

Tool pyzbar-mcp/scan_barcode
ARG image_paths=[./media/png]
OUT Barcode scan completed successfully. Found ISBN(s): 9780670856688. No further barcode scanning is needed -- proceeding directly to book lookup.

Step 2

Tool openlibrary_mcp/get_book_info
ARG isbn="9780670856688"
OUT Retrieved metadata from OpenLibrary: *The Discovery of Heaven* by Harry Mulisch; published by Viking in 1996, 730 pages. Query type confirmed as ISBN lookup.

Final Reply

Title: *The Discovery of Heaven*.
Author: Harry Mulisch.
ISBN-13: 9780670856688 (ISBN-10: 0670856681).
Publisher: Viking; Year: 1996; Pages: 730.
Book successfully recognized from the scanned barcode using pyzbar-mcp and validated via OpenLibrary metadata.

QR Attraction QR decoding to retrieve image, location inference with geocoding, and air-quality query for current AQI and guidance. See (14) QR Attraction.

Example: QR-linked Attraction and Air Quality Retrieval

IMG

Q Pull the webpage link hidden in the image, grab the main photo from that page, infer the location shown, turn that location into coordinates, check current air quality there, and report AQI, pollutant levels, and health advice.

Step 1

Tool pyzbar-mcp/scan_barcode
ARG image_paths=./media/png
OUT Barcode scan completed successfully. Decoded one QRCode linking to https://unsplash.com/photos/people-near-building-_K2uemedTGo. No further scanning required -- proceeding to image retrieval.

Continued on next page

Example: QR-linked Attraction and Air Quality Retrieval (cont.)

Step 2

Tool	linkimage-mcp/fetch_unsplash_image
ARG	url="https://unsplash.com/photos/people-near-building-_K2uemedTGo"
OUT	Downloaded Unsplash photo by Nick Fewings successfully. Image saved locally and accessible at ./images/_K2uemedTGo.jpg.

Step 3

Tool	google-maps/geocode
ARG	address="Natural History Museum, London"; language="en"
OUT	Geocoding completed -- location resolved to Natural History Museum, London. Coordinates: 51.496715, -0.1763672 (Cromwell Rd, South Kensington, London SW7 5BD, UK).

Step 4

Tool	google-air/current_conditions
ARG	lat=51.496715; lng=-0.1763672
OUT	Fetched current air-quality data for London (Natural History Museum area). AQI: 82 -- Excellent air quality. Dominant pollutant: NO_2.

Final Reply

```
QR decoded to Unsplash link.
Main image downloaded successfully (Nick Fewings).
Detected location: Natural History Museum, London (51.496715, -0.1763672).
Current Air Quality Index (UAQI): 82 -- Excellent. Dominant pollutant: NO2.
Health guidance: All individuals may continue outdoor activity safely; those with asthma or NO2
sensitivity should still avoid prolonged exposure to traffic-heavy roads. Ventilation and light
exercise remain safe.
```

Biology Chain Organism detection with privacy blur, Wikipedia-based ecological roles, and validation of a plausible food chain. See (16) Biology Chain.

Example: BiologyChain — Organism Listing, Privacy Blur, and Food-Web Check

IMG	
Q	Can you check this picture, note every organism, blur any people for privacy, figure out what each one eats and where it sits in the web, and tell me if they could make a valid food chain?

Step 1

Tool	dinox-mcp/detect-all-objects
ARG	imageFileUri=./media/png; includeDescription=false
OUT	Objects detected: wolf (1), person (1), fox (1), rabbit (1). Bboxes use {xmin, ymin, xmax, ymax}; coordinates describe exact positions and spatial relations.

Step 2

Tool	imagesorcery-mcp/blur
ARG	input_path=./media/png; areas=[{x1=120,y1=93,x2=528,y2=1330, ...}]; invert_areas=false; output_path=./media/blurred.png
OUT	Privacy blur applied to the person region. Output: ./media/blurred.png.
Tool	wiki/summary
ARG	title="Rabbit"
OUT	Rabbits are small herbivorous mammals (family Leporidae) feeding on grasses and herbs; primary consumers and common prey for foxes and wolves.
Tool	wiki/summary
ARG	title="Red fox"
OUT	Red fox (*Vulpes vulpes*) -- an omnivorous mesopredator eating rabbits, rodents, birds, insects, and fruits; trophic level ≈ 3.
Tool	wiki/summary
ARG	titles="Gray wolf"
OUT	Wolf (*Canis lupus*) -- apex carnivore hunting large and small mammals (including hares and occasionally foxes); trophic level ≈ 4.
Tool	wiki/summary
ARG	title="Human"

Continued on next page

Example: BiologyChain — Organism Listing, Privacy Blur, and Food-Web Check (cont.)

OUT	Humans (*Homo sapiens*) -- omnivorous primates with advanced cognition; top-level consumers but not part of the natural food chain here.
Tool	wiki/summary
ARG	title="Grass"
OUT	Grasses (Poaceae) -- photosynthetic primary producers forming the base of terrestrial food webs; trophic level 1.
Tool	wiki/summary (Tree)
ARG	title="Tree"
OUT	Trees -- perennial woody producers performing photosynthesis; provide structure and habitat within ecosystems.

Final Reply

```
Privacy-safe image:  ./media/blurred.png

Organisms noted
- Grass/green plants -- producers (photosynthesis).
- Trees -- producers.
- Rabbit -- herbivore; primary consumer.
- Fox -- omnivore/mesopredator; secondary consumer.
- Wolf -- apex carnivore; tertiary consumer.
- Person -- human omnivore (not part of wild chain).

Valid food chain?
Yes.  Examples:
1) Grass → Rabbit → Wolf (common).
2) Grass → Rabbit → Fox → Wolf (ecologically possible; wolves often kill foxes as competitors and may occasionally consume them).

Plants act as producers; rabbit as primary consumer; fox as secondary; wolf as tertiary/apex.  The organisms present can form a valid trophic sequence.
```

E Error Trajectory Distribution Analysis

We illustrate representative failure trajectories to analyze the characteristic error patterns of different multimodal large language models (MLLMs). Each case is presented following the same format.

GLM-4.5v In the *Fruit Cal* task, `glm-4.5v` exhibits an *Illegal calling format*: the reply is wrapped with non-JSON delimiters, violating the MCP schema for `name`/`arguments`. This malformed structure prevents any valid tool execution and halts progress.

Illegal calling format
Task: Fruit Cal
Model: glm-4.5v

IMG



Q If a woman ate only the fruits shown here for all three meals, would that cover minimum daily calories, and what brief long-term health concerns might that raise?

Final Reply

```
"<|begin_of_box|>""name":"food_mcp/get_nutrition"
"arguments":{"query":"1 watermelon"}"<|end_of_box|> ..."
Unconventional invocation format using "<|begin_of_box|> ...  <|end_of_box|>"
```

Qwen2.5-VL-72B-Instruct In the *Movies* task, `Qwen2.5-VL-72B-Instruct` repeatedly issues *Unknown Tool Invocation* to `ocr/perform` and `tmdb/*` endpoints absent from the registry, and later misuses a generic `server/tool` wrapper. Without tool discovery or correction, the trajectory culminates in an unsupported final claim.

Unknown Tool Invocation
Task: Movies
Model: Qwen2.5-VL-72B-Instruct

IMG



Q Can you pull the movie title from the image, look it up, and tell me where it's playing near New York?

Step 1 (Model attempted to call an MCP tool with non-existent tool name)

Tool ocr/perform
ARG {"input_data": [{"image": "./media/00020004.png", "language": "eng", "config": "-oem 3 -psm 6"}]}
OUT [Tool error] Unknown tool: ocr/perform

Step 2 (Parallel invalid tool calls with non-existent tool name)

Tool ocr/perform
ARG images=[{"path":"./media/png"}]
OUT [Tool error] Unknown tool: ocr/perform

Tool tmdb/search/movie
ARG query="Friday the 13th"
OUT [Tool error] Unknown tool: tmdb/search/movie

Tool tmdb/get_movie_details
ARG movie_id="REPLACE_WITH_ACTUAL_MOVIE_ID"
OUT [Tool error] Unknown tool: tmdb/get_movie_details

Tool tmdb/get_movie_showtimes
ARG movie_id="REPLACE_WITH_ACTUAL_MOVIE_ID"; region="US"; locale="en_US"; latitude=40.7128; longitude=-74.006
OUT [Tool error] Unknown tool: tmdb/get_movie_showtimes

Step 3 (Repeated the same invalid MCP call structure with no tool correction)

Tool ocr/perform
ARG {"input_data": {"image": "./media/png"}}
OUT [Tool error] Unknown tool: ocr/perform

Tool tmdb/search/movie
ARG query="Friday the 13th"
OUT [Tool error] Unknown tool: tmdb/search/movie

Tool tmdb/get_movie_details
ARG movie_id="12345"
OUT [Tool error] Unknown tool: tmdb/get_movie_details

Tool tmdb/get_movie_showtimes
ARG movie_id="12345"; region="NY"
OUT [Tool error] Unknown tool: tmdb/get_movie_showtimes

Step 4 (Used `server/tool` wrapper incorrectly for TMDB MCP call)

Tool server/tool
ARG {"tool": "tmdb/search_movies", "query": {"query": "Friday the 13th", "year": "2009"}}
OUT [Tool error] Unknown tool: server/tool

Final Reply

```
The movie "Friday the 13th" is not currently showing near New York.
You may want to check back later for updated showtimes.
```

InternVL-3.5 In the *Food Reddit, Wiki Excel* task, `InternVL-3.5` exhibits *Unknown Tool Invocation* with repeated calls to non-existent or incorrectly named tools (e.g., hallucinated tool endpoints for food nutrition, wiki, and Excel queries) and no corrective recovery. The failure mode mirrors the prior case, ending with a fabricated conclusion unsupported by executed tools.

Unknown Tool Invocation
Task: Food Reddit, Wiki Excel
Model: InternVL-3.5

IMG



Q From the foods in the image, can you list each name, find the hottest post about each, and note just its comment count (no repeated searches), then pull a short encyclopedia blurb and save everything to a spreadsheet with columns for Food Name and wiki Description?

Step 1 (Mixed errors: invalid invocation arguments and non-existent tools)

Tool	food_nutrition_mcp/get_food_nutrition
ARG	food="green beans"
OUT	Error executing tool food_nutrition_mcp/get_food_nutrition
Tool	wiki/search
ARG	query="green beans"; n=1
OUT	Top results: Green bean
Tool	excel/create_workbook
ARG	{}
OUT	Error executing tool create_workbook: Field required [filepath]
Tool	excel/create_worksheet
ARG	sheet_name="Food Info"
OUT	Error executing tool create_worksheet: Missing filepath argument
Tool	excel/write_data_to_excel
ARG	filepath="output.xlsx"; sheet_name="Food Info"; data=[["Food Name","Comment Count","Description"]]; start_cell="A1"
OUT	Error executing tool write_data_to_excel: Invalid filename -- must be absolute path when not in SSE mode

Step 2 (Parallel invalid tool calls with non-existent tool name)

Tool	food_nutrition_mcp/get_food_nutrition
ARG	food="green beans"
OUT	Error executing tool food_nutrition_mcp/get_food_nutrition
Tool	server/tool
ARG	tool="Reddit-MCP-Server/search_hot_posts"; subreddit="food"; limit=1
OUT	[Tool error] Unknown tool: server/tool
Tool	create_worksheet
ARG	filepath="/path/to/food_info.xlsx"; sheet_name="Food Info"
OUT	[Tool error] Unknown tool: create_worksheet
Tool	write_data_to_excel
ARG	filename="/path/to/food_info.xlsx"; sheet_name="Food Info"; data=[["Food Name","Comment Count","Description"]]
OUT	[Tool error] Unknown tool: write_data_to_excel

Step 3 (Mixed errors: invalid invocation arguments and non-existent tools)

Tool	food_nutrition_mcp/get_nutrition
ARG	query="green beans"
OUT	Error executing tool get_nutrition: 'measure'
Tool	Reddit-MCP-Server/search_hot_posts
ARG	subreddit="r/food"; limit=5
OUT	Error: Reddit API returned 400
Tool	excel/create_workbook
ARG	filepath="/path/to/output.xlsx"
OUT	Error: Permission denied at '/path'.
Tool	excel/create_worksheet
ARG	sheet_name="Food Info"; filepath="/path/to/output.xlsx"
OUT	Error: File not found -- '/path/to/output.xlsx'.
Tool	excel/write_data_to_excel
ARG	filepath="/path/to/output.xlsx"; sheet_name="Food Info"; data=[["Food Name","Comment Count","Description"]]; start_cell="A1"
OUT	Error: No such file or directory -- '/path/to/output.xlsx'.

Step 4 (Invalid invocation arguments)

Tool	Reddit-MCP-Server/search_hot_posts
ARG	subreddit="food"; limit=1; query="Green bean"
OUT	Top Reddit post found: 86 comments (r/food).
Tool	excel/create_worksheet

Continued on next page

Unknown Tool Invocation
Task: Food Reddit, Wiki Excel
Model: InternVL-3.5 (cont.)

ARG	`filepath="/path/to/output.xlsx"; sheet_name="Food Info"`
OUT	`Error: File not found -- '/path/to/output.xlsx'.`
Tool	`excel/write_data_to_excel`
ARG	`filepath="/path/to/output.xlsx"; sheet_name="Food Info"; data=[["Food Name","Comment Count","Description"],["Green bean","{comment_count}","{description}"]]; start_cell="A1"`
OUT	`Error: No such file or directory -- '/path/to/output.xlsx'.`

Final Reply

```
Here's the final compiled information for the food item:
Food Name: Green bean
Comment Count: 86
Description: Green beans are the young, unripe fruits of various cultivars of the common bean, also known as French beans, string beans, or snap beans. Popular worldwide for their mild flavor and versatility.
```

Claude-Sonnet-4-5 In the *Travel Comparison* task, `Claude-Sonnet-4-5` reveals a characteristic failure mode driven primarily by weak image understanding. The model misidentifies the right-hand landmark in the image, triggering a cascade of downstream geolocation errors. Finally, the trajectory ultimately produces a superficially coherent but unsupported final summary.

Incorrect image understanding
Task: Travel Comparison
Model: Claude-Sonnet-4-5

IMG



Q `Look at the side-by-side photo of two tourist spots, identify each landmark, get their coordinates, check current AQI and main pollutant levels for both, pick the healthier destination (if tied pick the first), find 2-3 nearby cinemas and 2-3 hotels for the chosen spot with approximate distances, and summarize.`

Step 1 (Irrelevant tool invocation. PowerPoint was not required for this task.)

Tool	`ppt/create_presentation`
ARG	`{"id": "tourist_comparison_presentation"}`
OUT	`Created new presentation with ID: tourist_comparison_presentation (0 slides).`

Step 2 (The image understanding of the model was incorrect. The right-hand landmark was not in Xining, China, but actually in Tokyo, Japan. This misidentification caused the subsequent geolocation and air-quality retrieval steps to fail.)

Tool	`google-maps/places_text_search`
ARG	`{"query": "White Pagoda Beijing Beihai Park", "maxResultCount": 3}`
OUT	`Top match → Beihai Park (39.9254° N, 116.3893° E).`
Tool	`google-maps/places_text_search`
ARG	`{"query": "Yakult building Xining China", "maxResultCount": 3}`
OUT	`Top match → Downtown Xining (36.6133° N, 101.8035° E).`

Step 3 (Model repeatedly attempted non-existent tools for air-quality retrieval.)

Tool	`weather_aqi/get_air_quality`
ARG	`{"lat": 39.9254, "lon": 116.3893}`
OUT	`50-150 (moderate to unhealthy)`
Tool	`weather_aqi/get_air_quality`
ARG	`{"lat": 36.6133, "lon": 101.8035}`
OUT	`30-80 (generally better)`

Final Reply

Continued on next page

Incorrect image understanding
Task: Travel Comparison
Model: Claude-Sonnet-4-5 (cont.)

Location Identification —
A: White Pagoda (Beihai Park, Beijing) (39.9254° N, 116.3892° E)
B: Yakult Building area (Xining, Qinghai Province) (36.6133° N, 101.8035° E)

Air Quality Comparison & Decision —
Live AQI queries failed because the model called unsupported tools. Using historical patterns:
Beijing = 50–150 (moderate to unhealthy), Xining = 30–80 (generally better).
Nevertheless, the model chose **Location A (Beihai Park)** for its tourist infrastructure.

Recommended Cinemas near Beihai Park (2–3 km):
1. Wangfujing CGV Cinema 2 km SE
2. Capital Cinema (Xidan) 2.5 km S
3. Broadway Cinematheque (MOMA) 3 km E

Recommended Hotels near Beihai Park (1–3 km):
1. Shichahai Shadow Art Performance Hotel 1 km N
2. Beijing Jianguo Hotel Qianmen 3 km S
3. Grand Hyatt Beijing 3.5 km SE

Overall verdict: Model achieved accurate landmark resolution but failed in air-quality retrieval due to non-existent MCP tools. Final summarization remained coherent and informative.

F PER-TASKS M^3-BENCH EVALUATION

Corresponding to Table 2, we present the evaluation results of each task described in Appendix B across all models, ordered by average score from lowest to highest.

Table 17: Sources for the model names in Table 2 and the task-level results.

Model	Source
GLM 4.5v	(Hong et al., 2025)
Qwen2.5-VL-72B	(Bai et al., 2025b)
InternVL 3.5	(Wang et al., 2025a)
Gemini 2.5 family	(Comanici et al., 2025)
Claude 4.5 Haiku	(Anthropic, 2025a)
Qwen3-VL-235B-A22B	(Bai et al., 2025a)
GPT-5 Nano	(OpenAI, 2025b)
L4-Scout-17B16E	(Meta AI, 2025)
Grok-4 Fast	(xAI, 2025b)
Claude 4.5 Sonnet	(Anthropic, 2025b)
GPT-5 Mini	(OpenAI, 2025a)
Grok 4 (0709)	(xAI, 2025a)
GPT-5	(OpenAI, 2025c)
Gemini 3.1 Pro	(Google DeepMind, 2026)

(0) Task: Items Price									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Claude 4.5 Haiku	0.000	0.000	0.000	0.000	0.000	0.000	0.386	0.247	0.079
Qwen2.5-VL-72B	0.000	0.000	0.000	0.000	0.000	0.000	0.303	0.388	0.086
Qwen3-VL-235B-A22B	0.055	0.056	0.063	0.051	0.050	0.050	0.335	0.430	0.134
Grok-4 Fast	0.200	0.100	0.190	0.067	0.000	0.200	0.501	0.378	0.204
InternVL 3.5	0.264	0.248	0.239	0.236	0.000	0.264	0.325	0.298	0.234
Gemini 2.5 Flash Lite	0.217	0.400	0.177	0.217	0.000	0.217	0.359	0.292	0.235
L4-Scout-17B16E	0.327	0.114	0.268	0.275	0.000	0.329	0.356	0.322	0.249
Gemini 2.5 Pro	0.360	0.189	0.250	0.360	0.000	0.360	0.487	0.427	0.304
Gemini 2.5 Flash	0.400	0.264	0.286	0.400	0.000	0.400	0.498	0.443	0.337
Claude 4.5 Sonnet	0.400	0.260	0.304	0.400	0.000	0.400	0.535	0.451	0.344
GPT-5 Mini	0.400	0.193	0.390	0.300	0.000	0.400	0.571	0.566	0.353
GPT-5	0.400	0.167	0.387	0.400	0.000	0.400	0.640	0.603	0.375
GPT-5 Nano	0.550	0.492	0.535	0.550	0.000	0.550	0.502	0.478	0.457
Grok 4 (0709)	0.700	0.207	0.496	0.700	0.000	0.700	0.606	0.530	0.492
GPT-5.5	0.967	0.227	0.902	0.464	0.000	0.967	0.809	0.950	0.661
Claude Opus 4.8	0.967	0.507	0.772	0.336	0.000	0.967	0.838	0.955	0.668
Gemini 3.1 Pro	1.000	0.353	0.864	0.583	0.000	1.000	0.738	0.833	0.671

Table 18: (0) Task: Items Price, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(1) Task: Attraction									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.167	0.067	0.135	0.111	0.000	0.167	0.285	0.130	0.133
Qwen2.5-VL-72B	0.133	0.031	0.118	0.067	0.000	0.133	0.303	0.388	0.147
Qwen3-VL-235B-A22B	0.174	0.080	0.170	0.113	0.052	0.177	0.341	0.424	0.195
L4-Scout-17B16E	0.333	0.049	0.249	0.333	0.000	0.333	0.367	0.335	0.250
Grok-4 Fast	0.278	0.403	0.257	0.278	0.000	0.278	0.501	0.378	0.296
InternVL 3.5	0.444	0.232	0.400	0.389	0.000	0.444	0.325	0.298	0.317
GPT-5 Mini	0.333	0.177	0.329	0.333	0.000	0.333	0.571	0.566	0.330
GPT-5 Nano	0.389	0.390	0.385	0.389	0.000	0.389	0.502	0.478	0.365
Gemini 2.5 Flash	0.500	0.200	0.462	0.361	0.000	0.500	0.498	0.443	0.371
Gemini 2.5 Flash Lite	0.500	0.500	0.402	0.500	0.000	0.500	0.359	0.292	0.382
Claude 4.5 Sonnet	0.489	0.443	0.430	0.506	0.000	0.399	0.518	0.433	0.402
GPT-5	0.444	0.272	0.417	0.444	0.000	0.444	0.640	0.603	0.408
Claude 4.5 Haiku	0.667	0.833	0.573	0.667	0.000	0.667	0.386	0.247	0.505
Grok 4 (0709)	0.611	0.526	0.565	0.611	0.000	0.611	0.606	0.530	0.508
Gemini 2.5 Pro	0.778	0.611	0.609	0.500	0.000	0.778	0.487	0.427	0.524
GPT-5.5	0.905	0.191	0.862	0.524	0.000	0.905	0.693	0.857	0.617
Claude Opus 4.8	0.857	0.750	0.758	0.690	0.000	0.857	0.950	0.834	0.712
Gemini 3.1 Pro	0.952	0.685	0.838	0.881	0.000	0.952	0.825	0.857	0.749

Table 19: (1) Task: Attraction, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(2) Task: Movies									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
InternVL 3.5	0.111	0.200	0.082	0.111	0.000	0.111	0.325	0.298	0.155
Qwen2.5-VL-72B	0.200	0.267	0.169	0.133	0.000	0.200	0.303	0.388	0.207
Qwen3-VL-235B-A22B	0.242	0.311	0.220	0.176	0.051	0.240	0.337	0.432	0.247
Claude 4.5 Haiku	0.333	0.500	0.244	0.333	0.250	0.006	0.386	0.247	0.287
Claude 4.5 Sonnet	0.250	0.417	0.196	0.250	0.125	0.083	0.535	0.451	0.288
L4-Scout-17B16E	0.333	0.398	0.258	0.333	0.000	0.333	0.367	0.335	0.295
GPT-5 Nano	0.278	0.306	0.263	0.278	0.000	0.278	0.502	0.478	0.298
GPT-5 Mini	0.389	0.344	0.374	0.389	0.000	0.389	0.571	0.566	0.378
Gemini 2.5 Flash	0.444	0.667	0.351	0.444	0.250	0.444	0.498	0.443	0.443
Gemini 2.5 Flash Lite	0.611	0.369	0.470	0.500	0.500	0.611	0.359	0.292	0.464
Gemini 2.5 Pro	0.556	0.522	0.401	0.556	0.333	0.556	0.487	0.427	0.480
Gemini 3.1 Pro	0.524	0.512	0.417	0.476	0.000	0.524	0.852	0.953	0.532
Grok-4 Fast	0.611	0.792	0.513	0.556	0.500	0.611	0.501	0.378	0.558
Claude Opus 4.8	0.619	0.583	0.477	0.619	0.000	0.619	0.804	0.929	0.581
Grok 4 (0709)	0.722	0.607	0.579	0.667	0.583	0.722	0.606	0.530	0.627
GPT-5	0.889	0.519	0.849	0.833	0.667	0.249	0.640	0.603	0.656
GPT-5.5	1.000	0.350	0.838	0.905	1.000	0.581	0.764	1.000	0.805

Table 20: (2) Task: Movies, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(3) Task: Vision Research									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Claude 4.5 Haiku	0.029	0.067	0.027	0.029	0.000	0.029	0.386	0.247	0.102
InternVL 3.5	0.204	0.137	0.182	0.176	0.018	0.204	0.325	0.298	0.193
L4-Scout-17B16E	0.211	0.093	0.198	0.211	0.029	0.211	0.367	0.335	0.207
GPT-5 Nano	0.167	0.053	0.153	0.139	0.000	0.167	0.502	0.478	0.208
GPT-5 Mini	0.167	0.082	0.159	0.167	0.000	0.167	0.471	0.466	0.210
GPT-5	0.167	0.115	0.159	0.167	0.000	0.167	0.640	0.603	0.252
Gemini 2.5 Flash Lite	0.167	0.793	0.155	0.167	0.000	0.167	0.359	0.292	0.263
Grok-4 Fast	0.290	0.258	0.267	0.269	0.061	0.290	0.501	0.378	0.289
Claude 4.5 Sonnet	0.291	0.258	0.270	0.291	0.079	0.291	0.535	0.451	0.308
Gemini 2.5 Flash	0.265	0.510	0.246	0.245	0.056	0.265	0.498	0.443	0.316
GPT-5.5	0.494	0.268	0.413	0.347	0.194	0.494	0.332	0.232	0.347
Grok 4 (0709)	0.363	0.538	0.318	0.363	0.112	0.363	0.606	0.530	0.399
Gemini 2.5 Pro	0.486	0.675	0.414	0.437	0.214	0.486	0.487	0.427	0.453
Gemini 3.1 Pro	0.686	0.480	0.538	0.565	0.426	0.686	0.289	0.095	0.471
Qwen2.5-VL-72B	0.600	0.500	0.550	0.600	0.250	0.600	0.303	0.388	0.474
Qwen3-VL-235B-A22B	0.617	0.520	0.579	0.625	0.293	0.625	0.341	0.427	0.501
Claude Opus 4.8	0.751	0.923	0.658	0.602	0.518	0.751	0.496	0.273	0.622

Table 21: (3) Task: Vision Research, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(4) Task: Camping									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.015	0.029	0.013	0.015	0.000	0.010	0.262	0.118	0.058
Gemini 2.5 Flash Lite	0.042	0.083	0.040	0.042	0.013	0.003	0.316	0.259	0.100
Qwen2.5-VL-72B	0.162	0.093	0.147	0.152	0.041	0.090	0.279	0.353	0.165
Qwen3-VL-235B-A22B	0.209	0.140	0.199	0.200	0.092	0.137	0.322	0.391	0.210
InternVL 3.5	0.254	0.207	0.223	0.254	0.131	0.099	0.319	0.292	0.222
Claude 4.5 Haiku	0.219	0.608	0.195	0.231	0.099	0.087	0.379	0.224	0.255
L4-Scout-17B16E	0.417	0.223	0.382	0.347	0.128	0.233	0.337	0.305	0.296
Gemini 2.5 Flash	0.341	0.362	0.315	0.258	0.178	0.217	0.496	0.449	0.327
Grok-4 Fast	0.390	0.455	0.363	0.345	0.229	0.232	0.461	0.344	0.352
GPT-5 Nano	0.433	0.181	0.427	0.403	0.238	0.269	0.502	0.478	0.366
Gemini 2.5 Pro	0.446	0.490	0.419	0.431	0.243	0.142	0.487	0.427	0.386
Claude 4.5 Sonnet	0.449	0.458	0.416	0.472	0.238	0.206	0.492	0.410	0.393
Grok 4 (0709)	0.520	0.391	0.486	0.499	0.369	0.347	0.606	0.530	0.469
GPT-5	0.562	0.382	0.547	0.488	0.356	0.326	0.616	0.594	0.484
GPT-5 Mini	0.583	0.374	0.573	0.562	0.402	0.342	0.571	0.566	0.497
Gemini 3.1 Pro	0.633	0.475	0.584	0.568	0.484	0.461	0.559	0.752	0.565
GPT-5.5	0.831	0.403	0.714	0.680	0.715	0.652	0.731	0.889	0.702
Claude Opus 4.8	0.887	0.972	0.781	0.748	0.839	0.863	0.836	0.966	0.862

Table 22: (4) Task: Camping, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(5) Task: Fish									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.018	0.091	0.013	0.018	0.000	0.018	0.285	0.130	0.072
InternVL 3.5	0.061	0.028	0.052	0.045	0.035	0.041	0.325	0.298	0.111
Qwen2.5-VL-72B	0.050	0.125	0.032	0.050	0.000	0.050	0.303	0.388	0.125
Claude 4.5 Haiku	0.160	0.137	0.139	0.160	0.100	0.021	0.386	0.247	0.169
Qwen3-VL-235B-A22B	0.098	0.171	0.092	0.099	0.052	0.099	0.340	0.427	0.173
Gemini 2.5 Flash Lite	0.179	0.094	0.154	0.161	0.051	0.123	0.359	0.292	0.176
GPT-5 Nano	0.164	0.035	0.152	0.164	0.010	0.127	0.502	0.478	0.204
Claude 4.5 Sonnet	0.317	0.220	0.276	0.317	0.097	0.043	0.535	0.451	0.282
L4-Scout-17B16E	0.400	0.135	0.339	0.367	0.210	0.198	0.367	0.335	0.294
Gemini 2.5 Flash	0.327	0.213	0.286	0.312	0.111	0.216	0.498	0.443	0.301
Grok-4 Fast	0.376	0.218	0.318	0.345	0.132	0.174	0.501	0.378	0.305
GPT-5 Mini	0.315	0.078	0.302	0.315	0.091	0.279	0.571	0.566	0.315
Gemini 2.5 Pro	0.533	0.299	0.418	0.418	0.260	0.464	0.487	0.427	0.413
GPT-5	0.552	0.134	0.508	0.506	0.162	0.327	0.640	0.603	0.429
Grok 4 (0709)	0.630	0.346	0.550	0.512	0.259	0.272	0.606	0.530	0.463
Gemini 3.1 Pro	1.000	0.627	0.867	0.909	0.917	0.736	0.759	0.450	0.783
GPT-5.5	1.000	0.415	0.901	0.891	0.799	0.584	0.916	0.807	0.789
Claude Opus 4.8	1.000	0.857	0.880	0.982	0.945	0.556	0.853	0.773	0.856

Table 23: (5) Task: Fish, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(6) Task: Drug									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.011	0.125	0.007	0.011	0.000	0.011	0.285	0.130	0.073
Gemini 2.5 Flash Lite	0.055	0.333	0.048	0.055	0.003	0.030	0.359	0.292	0.147
Qwen2.5-VL-72B	0.111	0.204	0.087	0.082	0.009	0.051	0.303	0.388	0.155
Gemini 2.5 Flash	0.106	0.189	0.095	0.055	0.003	0.106	0.498	0.443	0.187
Claude 4.5 Haiku	0.171	0.338	0.149	0.154	0.016	0.110	0.386	0.247	0.196
Qwen3-VL-235B-A22B	0.153	0.243	0.142	0.131	0.060	0.099	0.339	0.426	0.198
Grok-4 Fast	0.143	0.247	0.126	0.143	0.046	0.053	0.501	0.378	0.204
InternVL 3.5	0.208	0.479	0.171	0.157	0.028	0.010	0.325	0.298	0.210
Gemini 2.5 Pro	0.116	0.343	0.109	0.116	0.001	0.103	0.487	0.427	0.213
GPT-5 Nano	0.144	0.357	0.136	0.130	0.003	0.121	0.502	0.478	0.234
Grok 4 (0709)	0.155	0.258	0.147	0.135	0.003	0.133	0.606	0.530	0.246
GPT-5 Mini	0.204	0.213	0.196	0.137	0.026	0.177	0.571	0.566	0.261
Claude 4.5 Sonnet	0.276	0.422	0.250	0.276	0.045	0.234	0.535	0.451	0.311
L4-Scout-17B16E	0.439	0.594	0.367	0.199	0.082	0.207	0.367	0.335	0.324
GPT-5	0.382	0.440	0.361	0.351	0.101	0.289	0.640	0.603	0.396
Gemini 3.1 Pro	0.364	0.731	0.289	0.336	0.103	0.236	0.859	0.325	0.406
Claude Opus 4.8	0.439	0.688	0.350	0.425	0.204	0.372	0.922	0.305	0.463
GPT-5.5	0.696	0.560	0.609	0.351	0.214	0.330	0.637	0.385	0.473

Table 24: (6) Task: Drug, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(7) Task: Hazard									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Claude 4.5 Haiku	0.000	0.000	0.000	0.000	0.000	0.000	0.386	0.247	0.079
Gemini 2.5 Flash Lite	0.143	0.160	0.110	0.127	0.008	0.143	0.359	0.292	0.168
Qwen2.5-VL-72B	0.163	0.045	0.138	0.139	0.020	0.163	0.303	0.388	0.170
GPT-5 Nano	0.124	0.021	0.118	0.124	0.016	0.124	0.502	0.478	0.189
L4-Scout-17B16E	0.205	0.123	0.170	0.185	0.018	0.164	0.367	0.335	0.196
Qwen3-VL-235B-A22B	0.208	0.097	0.189	0.182	0.069	0.204	0.341	0.425	0.217
InternVL 3.5	0.278	0.239	0.219	0.259	0.091	0.049	0.325	0.298	0.220
GPT-5 Mini	0.148	0.077	0.139	0.130	0.040	0.101	0.571	0.566	0.221
Grok-4 Fast	0.421	0.410	0.356	0.402	0.141	0.263	0.501	0.378	0.359
Gemini 2.5 Flash	0.519	0.333	0.436	0.500	0.310	0.259	0.498	0.443	0.412
Grok 4 (0709)	0.608	0.280	0.499	0.571	0.245	0.364	0.606	0.530	0.463
Claude 4.5 Sonnet	0.656	0.321	0.556	0.638	0.386	0.193	0.535	0.451	0.467
Gemini 2.5 Pro	0.714	0.442	0.565	0.698	0.395	0.192	0.487	0.427	0.490
GPT-5	0.802	0.251	0.695	0.733	0.317	0.328	0.640	0.603	0.546
GPT-5.5	0.750	0.571	0.625	0.600	0.395	0.613	0.598	0.700	0.606
Gemini 3.1 Pro	0.767	0.508	0.600	0.650	0.490	0.618	0.500	0.720	0.607
Claude Opus 4.8	0.850	0.633	0.689	0.683	0.619	0.786	0.645	0.820	0.716

Table 25: (7) Task: Hazard, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(8) Task: Food Reddit Wiki Excel									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Claude 4.5 Haiku	0.006	0.017	0.005	0.006	0.000	0.006	0.386	0.247	0.084
Qwen2.5-VL-72B	0.171	0.032	0.156	0.135	0.057	0.103	0.303	0.388	0.168
Gemini 2.5 Flash Lite	0.163	0.168	0.143	0.128	0.029	0.163	0.359	0.292	0.181
Qwen3-VL-235B-A22B	0.207	0.078	0.204	0.180	0.108	0.148	0.345	0.429	0.212
GPT-5 Nano	0.378	0.104	0.373	0.206	0.000	0.378	0.502	0.478	0.302
L4-Scout-17B16E	0.476	0.488	0.446	0.311	0.000	0.476	0.367	0.335	0.362
InternVL 3.5	0.525	0.556	0.476	0.351	0.000	0.525	0.325	0.298	0.382
GPT-5 Mini	0.562	0.322	0.553	0.290	0.054	0.539	0.571	0.566	0.432
Gemini 2.5 Flash	0.599	0.496	0.544	0.298	0.229	0.589	0.498	0.443	0.462
Grok-4 Fast	0.636	0.460	0.571	0.378	0.290	0.549	0.501	0.378	0.470
Claude 4.5 Sonnet	0.634	0.674	0.574	0.634	0.229	0.448	0.535	0.451	0.522
Grok 4 (0709)	0.685	0.639	0.635	0.389	0.113	0.685	0.606	0.530	0.535
Claude Opus 4.8	0.623	0.796	0.565	0.329	0.335	0.623	0.550	0.472	0.536
GPT-5	0.765	0.416	0.708	0.379	0.175	0.765	0.640	0.603	0.556
Gemini 2.5 Pro	0.752	0.622	0.645	0.656	0.386	0.737	0.487	0.427	0.589
GPT-5.5	0.829	0.700	0.744	0.376	0.504	0.773	0.336	0.476	0.592
Gemini 3.1 Pro	0.806	0.829	0.732	0.428	0.496	0.779	0.533	0.652	0.657

Table 26: (8) Task: Food Reddit Wiki Excel, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(9) Task: MedExcel									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.004	0.009	0.003	0.005	0.000	0.005	0.266	0.121	0.052
Gemini 2.5 Flash Lite	0.036	0.013	0.027	0.036	0.000	0.036	0.316	0.259	0.090
InternVL 3.5	0.071	0.029	0.049	0.071	0.000	0.071	0.319	0.292	0.113
GPT-5 Nano	0.050	0.021	0.043	0.050	0.000	0.020	0.402	0.378	0.121
Claude 4.5 Haiku	0.067	0.189	0.052	0.079	0.006	0.044	0.385	0.229	0.131
GPT-5 Mini	0.071	0.041	0.059	0.071	0.000	0.071	0.471	0.466	0.156
Grok-4 Fast	0.118	0.141	0.097	0.118	0.013	0.117	0.468	0.351	0.178
Gemini 2.5 Flash	0.107	0.086	0.085	0.107	0.000	0.107	0.496	0.449	0.180
Grok 4 (0709)	0.071	0.057	0.060	0.071	0.000	0.071	0.606	0.530	0.183
Claude 4.5 Sonnet	0.136	0.142	0.111	0.161	0.013	0.104	0.500	0.419	0.198
Gemini 2.5 Pro	0.143	0.114	0.115	0.143	0.024	0.143	0.487	0.427	0.200
GPT-5	0.179	0.095	0.150	0.179	0.024	0.107	0.616	0.594	0.243
Qwen2.5-VL-72B	0.250	0.351	0.204	0.250	0.033	0.250	0.303	0.388	0.254
L4-Scout-17B16E	0.333	0.104	0.283	0.333	0.056	0.333	0.385	0.354	0.273
Qwen3-VL-235B-A22B	0.286	0.389	0.254	0.290	0.083	0.289	0.339	0.416	0.291
GPT-5.5	0.643	0.429	0.528	0.643	0.381	0.643	0.368	0.429	0.508
Gemini 3.1 Pro	0.821	0.907	0.618	0.821	0.643	0.789	0.564	0.554	0.715
Claude Opus 4.8	0.857	0.611	0.656	0.857	0.714	0.857	0.711	0.679	0.743

Table 27: (9) Task: MedExcel, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(10) Task: Barcode Lookup									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Gemini 2.5 Flash Lite	0.000	0.000	0.000	0.000	0.000	0.000	0.359	0.292	0.081
Grok-4 Fast	0.083	0.033	0.083	0.083	0.000	0.083	0.501	0.378	0.156
Qwen2.5-VL-72B	0.200	0.225	0.187	0.200	0.000	0.200	0.303	0.388	0.213
Claude 4.5 Haiku	0.250	0.417	0.230	0.250	0.000	0.250	0.386	0.247	0.254
Qwen3-VL-235B-A22B	0.245	0.270	0.240	0.245	0.050	0.237	0.338	0.422	0.260
GPT-5 Mini	0.333	0.417	0.333	0.333	0.000	0.333	0.571	0.566	0.361
InternVL 3.5	0.417	0.750	0.384	0.417	0.000	0.417	0.325	0.298	0.376
GPT-5 Nano	0.500	0.667	0.484	0.500	0.000	0.500	0.502	0.478	0.454
L4-Scout-17B16E	0.667	0.756	0.593	0.667	0.333	0.503	0.367	0.335	0.528
Claude 4.5 Sonnet	0.667	0.625	0.625	0.583	0.167	0.667	0.535	0.451	0.540
GPT-5	0.917	0.806	0.880	0.833	0.667	0.752	0.640	0.603	0.762
Gemini 2.5 Flash	0.917	1.000	0.878	0.833	0.667	0.917	0.498	0.443	0.769
Grok 4 (0709)	0.917	0.861	0.826	0.833	0.667	0.917	0.606	0.530	0.770
Gemini 2.5 Pro	1.000	1.000	0.962	0.917	0.833	1.000	0.487	0.427	0.828
Claude Opus 4.8	1.000	1.000	0.849	1.000	1.000	0.674	0.979	1.000	0.938
GPT-5.5	1.000	0.750	0.840	1.000	1.000	1.000	0.979	1.000	0.946
Gemini 3.1 Pro	1.000	1.000	0.834	1.000	1.000	1.000	1.000	1.000	0.979

Table 28: (10) Task: Barcode Lookup, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(11) Task: Biology Cut									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
Gemini 2.5 Flash Lite	0.000	0.000	0.000	0.000	0.000	0.000	0.359	0.292	0.081
InternVL 3.5	0.024	0.056	0.024	0.024	0.000	0.024	0.325	0.298	0.097
GLM 4.5v	0.125	0.028	0.117	0.125	0.000	0.125	0.285	0.130	0.117
GPT-5 Nano	0.000	0.000	0.000	0.000	0.000	0.000	0.502	0.478	0.123
GPT-5 Mini	0.000	0.000	0.000	0.000	0.000	0.000	0.571	0.566	0.142
Qwen2.5-VL-72B	0.400	0.118	0.375	0.200	0.000	0.400	0.303	0.388	0.273
Claude 4.5 Haiku	0.318	0.443	0.304	0.318	0.000	0.318	0.386	0.247	0.292
Qwen3-VL-235B-A22B	0.432	0.166	0.406	0.238	0.050	0.431	0.340	0.434	0.312
L4-Scout-17B16E	0.389	0.400	0.362	0.389	0.000	0.389	0.367	0.335	0.329
Gemini 2.5 Pro	0.425	0.345	0.390	0.363	0.167	0.425	0.487	0.427	0.379
Grok-4 Fast	0.743	0.262	0.698	0.391	0.000	0.743	0.501	0.378	0.464
Grok 4 (0709)	0.476	0.735	0.429	0.476	0.000	0.476	0.606	0.530	0.466
Claude 4.5 Sonnet	0.693	0.538	0.665	0.693	0.167	0.693	0.535	0.451	0.554
Gemini 2.5 Flash	0.575	0.739	0.549	0.575	0.500	0.575	0.498	0.443	0.557
GPT-5.5	0.785	0.976	0.563	0.713	0.000	0.785	0.887	0.583	0.662
Gemini 3.1 Pro	0.808	0.576	0.592	0.545	0.167	0.808	0.950	0.917	0.670
Claude Opus 4.8	0.818	0.893	0.612	0.746	0.167	0.818	0.925	0.822	0.725
GPT-5	0.958	0.939	0.919	0.958	0.833	0.958	0.640	0.603	0.851

Table 29: (11) Task: Biology Cut, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(12) Task: Fruit Cal									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.005	0.000	0.282	0.128	0.052
InternVL 3.5	0.071	0.034	0.045	0.069	0.007	0.069	0.324	0.298	0.115
Qwen2.5-VL-72B	0.113	0.059	0.089	0.110	0.011	0.110	0.301	0.388	0.147
Grok 4 (0709)	0.095	0.038	0.084	0.093	0.009	0.093	0.604	0.530	0.193
Qwen3-VL-235B-A22B	0.155	0.108	0.143	0.155	0.061	0.155	0.334	0.430	0.194
Claude 4.5 Haiku	0.207	0.268	0.167	0.202	0.020	0.202	0.381	0.247	0.212
L4-Scout-17B16E	0.225	0.127	0.194	0.219	0.022	0.220	0.362	0.335	0.213
Grok-4 Fast	0.164	0.271	0.148	0.160	0.016	0.160	0.497	0.378	0.224
GPT-5 Mini	0.164	0.160	0.157	0.160	0.016	0.160	0.567	0.566	0.244
GPT-5 Nano	0.307	0.082	0.290	0.172	0.017	0.303	0.498	0.478	0.268
Claude 4.5 Sonnet	0.293	0.476	0.272	0.286	0.029	0.286	0.528	0.451	0.328
Gemini 2.5 Pro	0.360	0.462	0.307	0.309	0.031	0.352	0.480	0.427	0.341
Gemini 2.5 Flash Lite	0.488	0.256	0.430	0.441	0.045	0.477	0.348	0.292	0.347
Gemini 2.5 Flash	0.460	0.952	0.416	0.448	0.046	0.449	0.487	0.443	0.463
GPT-5	0.481	0.782	0.427	0.400	0.041	0.471	0.630	0.603	0.479
Claude Opus 4.8	0.657	0.392	0.580	0.332	0.000	0.657	0.725	0.594	0.492
GPT-5.5	0.962	0.253	0.797	0.307	0.000	0.962	0.521	0.690	0.562
Gemini 3.1 Pro	0.905	0.396	0.793	0.417	0.000	0.905	0.539	0.691	0.581

Table 30: (12) Task: Fruit Cal, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(13) Task: Medical									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Qwen2.5-VL-72B	0.036	0.031	0.035	0.036	0.000	0.036	0.303	0.388	0.108
Qwen3-VL-235B-A22B	0.082	0.076	0.093	0.087	0.053	0.087	0.338	0.428	0.155
Claude 4.5 Haiku	0.152	0.472	0.119	0.152	0.026	0.091	0.386	0.247	0.206
InternVL 3.5	0.199	0.361	0.147	0.158	0.023	0.161	0.325	0.298	0.209
Gemini 2.5 Flash Lite	0.265	0.242	0.227	0.223	0.053	0.189	0.359	0.292	0.231
Gemini 2.5 Flash	0.215	0.444	0.198	0.215	0.053	0.215	0.498	0.443	0.285
L4-Scout-17B16E	0.396	0.318	0.332	0.233	0.108	0.279	0.367	0.335	0.296
Grok-4 Fast	0.223	0.647	0.201	0.223	0.093	0.223	0.501	0.378	0.311
Gemini 2.5 Pro	0.332	0.506	0.296	0.332	0.095	0.275	0.487	0.427	0.344
Claude 4.5 Sonnet	0.350	0.534	0.314	0.350	0.133	0.181	0.535	0.451	0.356
GPT-5	0.462	0.382	0.402	0.414	0.275	0.170	0.640	0.603	0.418
GPT-5 Nano	0.528	0.317	0.515	0.463	0.153	0.447	0.502	0.478	0.425
GPT-5 Mini	0.636	0.506	0.620	0.550	0.407	0.254	0.571	0.566	0.514
Grok 4 (0709)	0.568	0.466	0.506	0.544	0.418	0.492	0.606	0.530	0.516
Claude Opus 4.8	0.682	0.845	0.619	0.431	0.390	0.682	0.863	0.565	0.634
Gemini 3.1 Pro	0.813	0.828	0.710	0.644	0.645	0.686	0.725	0.634	0.711
GPT-5.5	1.000	0.465	0.933	0.588	0.925	0.750	0.808	0.715	0.773

Table 31: (13) Task: Medical, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(14) Task: QR Attraction									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.019	0.034	0.016	0.020	0.000	0.023	0.270	0.122	0.063
GPT-5 Nano	0.136	0.050	0.136	0.136	0.000	0.136	0.502	0.478	0.197
InternVL 3.5	0.214	0.236	0.168	0.214	0.000	0.214	0.319	0.292	0.207
Qwen2.5-VL-72B	0.280	0.302	0.220	0.280	0.022	0.280	0.303	0.388	0.259
Qwen3-VL-235B-A22B	0.321	0.345	0.273	0.316	0.071	0.313	0.334	0.430	0.298
Claude 4.5 Haiku	0.285	0.717	0.237	0.322	0.118	0.190	0.390	0.231	0.311
GPT-5 Mini	0.364	0.202	0.363	0.364	0.103	0.364	0.571	0.566	0.362
Claude 4.5 Sonnet	0.421	0.322	0.353	0.421	0.190	0.336	0.535	0.451	0.379
L4-Scout-17B16E	0.571	0.204	0.479	0.543	0.221	0.442	0.385	0.354	0.400
Gemini 2.5 Flash Lite	0.493	0.612	0.420	0.493	0.262	0.411	0.316	0.259	0.408
GPT-5	0.571	0.405	0.516	0.543	0.268	0.542	0.616	0.594	0.507
Claude Opus 4.8	0.679	0.519	0.582	0.679	0.452	0.641	0.514	0.643	0.589
Gemini 2.5 Flash	0.757	0.491	0.672	0.729	0.556	0.757	0.496	0.449	0.613
Grok-4 Fast	0.757	0.738	0.671	0.729	0.516	0.757	0.501	0.378	0.631
Gemini 2.5 Pro	0.829	0.485	0.745	0.800	0.651	0.794	0.487	0.427	0.652
Grok 4 (0709)	0.829	0.686	0.748	0.800	0.651	0.829	0.606	0.530	0.710
Gemini 3.1 Pro	0.929	0.831	0.795	0.929	0.857	0.929	0.714	0.607	0.824
GPT-5.5	1.000	0.608	0.871	1.000	1.000	1.000	0.721	0.643	0.855

Table 32: (14) Task: QR Attraction, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(15) Task: Routine									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.285	0.130	0.052
Gemini 2.5 Flash Lite	0.081	0.327	0.066	0.081	0.000	0.081	0.316	0.259	0.152
L4-Scout-17B16E	0.115	0.238	0.098	0.115	0.000	0.115	0.385	0.354	0.177
Qwen2.5-VL-72B	0.162	0.162	0.132	0.162	0.031	0.106	0.303	0.388	0.181
Qwen3-VL-235B-A22B	0.205	0.207	0.186	0.205	0.079	0.153	0.342	0.422	0.225
InternVL 3.5	0.239	0.417	0.195	0.194	0.029	0.191	0.319	0.292	0.234
Grok-4 Fast	0.105	0.714	0.086	0.105	0.000	0.105	0.501	0.378	0.249
Claude 4.5 Sonnet	0.190	0.311	0.168	0.190	0.029	0.142	0.535	0.451	0.252
Claude 4.5 Haiku	0.258	0.732	0.226	0.258	0.143	0.131	0.386	0.247	0.297
GPT-5 Nano	0.452	0.202	0.432	0.343	0.201	0.416	0.502	0.478	0.378
Gemini 2.5 Flash	0.475	0.566	0.419	0.261	0.171	0.475	0.496	0.449	0.414
Grok 4 (0709)	0.341	0.761	0.297	0.341	0.143	0.341	0.606	0.530	0.420
Gemini 2.5 Pro	0.447	0.929	0.396	0.447	0.221	0.369	0.487	0.427	0.465
GPT-5 Mini	0.635	0.344	0.614	0.537	0.370	0.446	0.571	0.566	0.510
GPT-5	0.613	0.422	0.565	0.359	0.462	0.538	0.616	0.594	0.521
Claude Opus 4.8	0.684	0.595	0.622	0.413	0.411	0.635	0.875	0.779	0.627
GPT-5.5	0.816	0.363	0.781	0.684	0.649	0.636	0.775	0.991	0.712
Gemini 3.1 Pro	0.821	0.668	0.729	0.612	0.601	0.778	0.768	0.842	0.727

Table 33: (15) Task: Routine, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(16) Task: Biology Chain									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.011	0.050	0.011	0.011	0.000	0.011	0.285	0.130	0.064
Qwen2.5-VL-72B	0.117	0.051	0.091	0.117	0.015	0.051	0.303	0.388	0.142
GPT-5 Nano	0.049	0.063	0.048	0.049	0.000	0.049	0.502	0.478	0.155
Gemini 2.5 Flash Lite	0.150	0.202	0.130	0.150	0.035	0.026	0.316	0.259	0.158
InternVL 3.5	0.090	0.375	0.068	0.090	0.009	0.058	0.319	0.292	0.163
GPT-5 Mini	0.071	0.031	0.071	0.036	0.000	0.071	0.571	0.566	0.177
Qwen3-VL-235B-A22B	0.159	0.099	0.148	0.162	0.066	0.097	0.335	0.427	0.187
L4-Scout-17B16E	0.318	0.367	0.268	0.293	0.098	0.183	0.385	0.354	0.283
Claude 4.5 Haiku	0.375	0.683	0.294	0.375	0.156	0.150	0.386	0.247	0.333
Grok-4 Fast	0.329	0.681	0.289	0.294	0.051	0.214	0.501	0.378	0.342
Gemini 2.5 Pro	0.360	0.651	0.287	0.335	0.104	0.290	0.487	0.427	0.368
Gemini 2.5 Flash	0.457	0.486	0.363	0.363	0.171	0.321	0.496	0.449	0.388
Claude 4.5 Sonnet	0.454	0.547	0.381	0.417	0.189	0.158	0.535	0.451	0.391
Grok 4 (0709)	0.411	0.663	0.334	0.380	0.155	0.342	0.606	0.530	0.428
Claude Opus 4.8	0.537	0.835	0.467	0.425	0.086	0.475	0.775	0.539	0.517
GPT-5.5	0.861	0.389	0.722	0.230	0.000	0.861	0.668	0.548	0.535
Gemini 3.1 Pro	0.770	0.642	0.671	0.317	0.114	0.770	0.643	0.513	0.555
GPT-5	0.810	0.284	0.677	0.633	0.466	0.430	0.616	0.594	0.564

Table 34: (16) Task: Biology Chain, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(17) Task: Generate PowerPoint									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.024	0.026	0.020	0.024	0.000	0.023	0.270	0.122	0.064
Qwen2.5-VL-72B	0.200	0.057	0.164	0.200	0.083	0.103	0.303	0.388	0.187
Gemini 2.5 Flash Lite	0.210	0.083	0.176	0.210	0.048	0.210	0.316	0.259	0.189
GPT-5 Nano	0.136	0.054	0.134	0.136	0.009	0.086	0.502	0.478	0.192
Qwen3-VL-235B-A22B	0.239	0.102	0.211	0.240	0.130	0.152	0.342	0.424	0.231
GPT-5 Mini	0.200	0.094	0.196	0.200	0.026	0.159	0.571	0.566	0.252
Claude 4.5 Haiku	0.350	0.536	0.306	0.384	0.186	0.193	0.390	0.231	0.322
Claude 4.5 Sonnet	0.333	0.365	0.310	0.333	0.071	0.333	0.535	0.451	0.342
GPT-5.5	0.467	0.340	0.375	0.406	0.170	0.354	0.242	0.492	0.356
Claude Opus 4.8	0.457	0.410	0.404	0.337	0.159	0.311	0.485	0.346	0.364
InternVL 3.5	0.552	0.326	0.478	0.524	0.328	0.365	0.319	0.292	0.398
Grok-4 Fast	0.467	0.234	0.420	0.438	0.369	0.467	0.501	0.378	0.409
Gemini 3.1 Pro	0.444	0.368	0.382	0.394	0.170	0.363	0.585	0.737	0.430
L4-Scout-17B16E	0.867	0.191	0.788	0.814	0.423	0.830	0.385	0.354	0.582
Grok 4 (0709)	0.771	0.422	0.706	0.748	0.655	0.490	0.606	0.530	0.616
Gemini 2.5 Flash	0.857	0.261	0.787	0.833	0.857	0.857	0.496	0.449	0.675
GPT-5	0.971	0.255	0.938	0.890	0.607	0.765	0.616	0.594	0.705
Gemini 2.5 Pro	0.924	0.567	0.840	0.924	0.833	0.812	0.487	0.427	0.727

Table 35: (17) Task: Generate PowerPoint, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(18) Task: Travel Comparison									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.023	0.017	0.020	0.020	0.000	0.024	0.270	0.122	0.062
Qwen2.5-VL-72B	0.111	0.026	0.108	0.111	0.000	0.111	0.303	0.388	0.145
Qwen3-VL-235B-A22B	0.157	0.076	0.163	0.156	0.053	0.158	0.342	0.426	0.190
Claude 4.5 Haiku	0.348	0.359	0.297	0.317	0.199	0.200	0.390	0.231	0.293
InternVL 3.5	0.476	0.224	0.400	0.381	0.429	0.476	0.319	0.292	0.375
Gemini 2.5 Pro	0.571	0.088	0.509	0.429	0.571	0.356	0.487	0.427	0.430
L4-Scout-17B16E	0.714	0.171	0.650	0.476	0.071	0.714	0.385	0.354	0.442
GPT-5 Nano	0.643	0.265	0.619	0.562	0.149	0.421	0.502	0.478	0.455
Claude 4.5 Sonnet	0.619	0.417	0.532	0.619	0.143	0.485	0.535	0.451	0.475
Grok 4 (0709)	0.619	0.099	0.543	0.524	0.571	0.416	0.606	0.530	0.488
Gemini 2.5 Flash Lite	0.714	0.388	0.601	0.619	0.643	0.386	0.316	0.259	0.491
Grok-4 Fast	0.667	0.446	0.601	0.571	0.429	0.667	0.501	0.378	0.532
GPT-5 Mini	0.714	0.203	0.691	0.714	0.438	0.481	0.571	0.566	0.547
GPT-5	1.000	0.119	0.938	0.810	0.857	0.573	0.616	0.594	0.688
Gemini 2.5 Flash	1.000	0.239	0.897	0.667	1.000	1.000	0.496	0.449	0.718
GPT-5.5	0.914	0.485	0.871	0.805	0.804	0.914	0.743	0.624	0.770
Claude Opus 4.8	0.952	0.774	0.866	0.624	0.905	0.879	0.725	0.677	0.800
Gemini 3.1 Pro	1.000	0.741	0.889	1.000	1.000	1.000	0.682	0.715	0.878

Table 36: (18) Task: Travel Comparison, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(19) Task: Heatmap									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.016	0.028	0.014	0.017	0.000	0.017	0.269	0.122	0.060
Claude 4.5 Haiku	0.040	0.200	0.037	0.040	0.000	0.040	0.386	0.247	0.124
Qwen2.5-VL-72B	0.172	0.090	0.158	0.177	0.038	0.146	0.286	0.364	0.179
Grok-4 Fast	0.100	0.219	0.088	0.100	0.038	0.072	0.501	0.378	0.187
InternVL 3.5	0.225	0.280	0.214	0.225	0.038	0.194	0.319	0.292	0.223
Qwen3-VL-235B-A22B	0.214	0.141	0.211	0.214	0.086	0.189	0.328	0.411	0.224
Gemini 2.5 Flash Lite	0.325	0.848	0.311	0.325	0.088	0.170	0.316	0.259	0.330
Claude 4.5 Sonnet	0.360	0.269	0.346	0.360	0.140	0.268	0.535	0.451	0.341
L4-Scout-17B16E	0.650	0.253	0.605	0.650	0.292	0.383	0.385	0.354	0.447
Gemini 2.5 Flash	0.500	0.488	0.479	0.500	0.188	0.500	0.496	0.449	0.450
GPT-5 Mini	0.619	0.127	0.586	0.524	0.429	0.524	0.571	0.566	0.493
GPT-5 Nano	0.714	0.123	0.688	0.667	0.571	0.524	0.502	0.478	0.533
Gemini 2.5 Pro	0.625	0.674	0.594	0.625	0.338	0.568	0.487	0.427	0.542
Grok 4 (0709)	0.700	0.520	0.687	0.700	0.525	0.657	0.606	0.530	0.616
GPT-5	0.750	0.653	0.710	0.750	0.538	0.683	0.616	0.594	0.662
GPT-5.5	1.000	0.381	0.977	0.905	1.000	0.511	0.693	0.917	0.798
Gemini 3.1 Pro	1.000	0.456	0.946	0.905	1.000	0.873	0.632	1.000	0.851
Claude Opus 4.8	1.000	0.779	0.920	0.667	1.000	1.000	0.836	1.000	0.900

Table 37: (19) Task: Heatmap, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(20) Task: Museum									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.014	0.034	0.012	0.010	0.000	0.014	0.262	0.118	0.058
Qwen2.5-VL-72B	0.150	0.109	0.138	0.101	0.012	0.124	0.279	0.353	0.158
InternVL 3.5	0.200	0.217	0.198	0.163	0.000	0.200	0.319	0.292	0.199
Qwen3-VL-235B-A22B	0.191	0.155	0.191	0.148	0.063	0.168	0.320	0.398	0.203
Claude 4.5 Haiku	0.203	0.714	0.183	0.153	0.028	0.120	0.379	0.224	0.250
L4-Scout-17B16E	0.385	0.262	0.359	0.231	0.036	0.321	0.337	0.305	0.279
Gemini 2.5 Flash Lite	0.438	0.334	0.415	0.275	0.064	0.295	0.316	0.259	0.299
Grok-4 Fast	0.360	0.534	0.341	0.229	0.064	0.320	0.461	0.344	0.332
Gemini 2.5 Pro	0.325	0.692	0.306	0.238	0.048	0.193	0.487	0.427	0.340
Claude 4.5 Sonnet	0.414	0.537	0.391	0.314	0.067	0.284	0.492	0.410	0.364
Gemini 2.5 Flash	0.375	0.604	0.359	0.267	0.076	0.350	0.496	0.449	0.372
GPT-5 Nano	0.475	0.229	0.465	0.475	0.150	0.425	0.502	0.478	0.400
Grok 4 (0709)	0.438	0.467	0.417	0.363	0.105	0.438	0.606	0.530	0.420
GPT-5 Mini	0.550	0.249	0.543	0.550	0.272	0.466	0.571	0.566	0.471
GPT-5	0.638	0.296	0.607	0.388	0.176	0.548	0.616	0.594	0.483
Gemini 3.1 Pro	0.800	0.655	0.721	0.800	0.650	0.800	0.506	0.625	0.695
GPT-5.5	0.925	0.792	0.839	0.925	0.850	0.925	0.594	0.825	0.834
Claude Opus 4.8	0.925	0.863	0.893	0.925	0.850	0.925	0.775	0.825	0.873

Table 38: (20) Task: Museum, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(21) Task: Research Station									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.086	0.067	0.084	0.080	0.032	0.079	0.325	0.213	0.118
Gemini 2.5 Flash Lite	0.195	0.113	0.191	0.172	0.057	0.192	0.330	0.325	0.196
Claude 4.5 Haiku	0.199	0.299	0.177	0.217	0.079	0.138	0.376	0.225	0.210
InternVL 3.5	0.263	0.200	0.228	0.252	0.114	0.198	0.332	0.312	0.240
GPT-5 Nano	0.212	0.249	0.207	0.222	0.093	0.125	0.459	0.463	0.256
Claude 4.5 Sonnet	0.270	0.315	0.259	0.268	0.065	0.264	0.505	0.415	0.298
Qwen2.5-VL-72B	0.331	0.234	0.314	0.341	0.094	0.286	0.372	0.437	0.300
Qwen3-VL-235B-A22B	0.331	0.234	0.314	0.341	0.094	0.286	0.372	0.437	0.300
GPT-5 Mini	0.300	0.177	0.290	0.300	0.167	0.258	0.556	0.550	0.325
Grok-4 Fast	0.395	0.393	0.379	0.381	0.167	0.395	0.519	0.427	0.376
GPT-5.5	0.438	0.521	0.410	0.438	0.186	0.438	0.200	0.406	0.379
L4-Scout-17B16E	0.568	0.204	0.529	0.545	0.242	0.547	0.407	0.392	0.436
Gemini 2.5 Flash	0.514	0.343	0.491	0.498	0.335	0.514	0.523	0.479	0.461
Claude Opus 4.8	0.550	0.743	0.521	0.412	0.252	0.508	0.300	0.417	0.463
Grok 4 (0709)	0.514	0.368	0.490	0.496	0.311	0.425	0.625	0.615	0.480
Gemini 2.5 Pro	0.582	0.423	0.555	0.524	0.367	0.545	0.541	0.507	0.502
GPT-5	0.678	0.301	0.649	0.613	0.446	0.449	0.644	0.615	0.561
Gemini 3.1 Pro	0.762	0.625	0.720	0.669	0.468	0.703	0.258	0.525	0.591

Table 39: (21) Task: Research Station, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(22) Task: Hugeicons									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.014	0.036	0.010	0.013	0.000	0.016	0.262	0.118	0.059
Gemini 2.5 Flash Lite	0.113	0.256	0.096	0.113	0.000	0.113	0.316	0.259	0.158
Qwen2.5-VL-72B	0.149	0.115	0.114	0.137	0.040	0.135	0.279	0.353	0.165
GPT-5.5	0.238	0.293	0.221	0.095	0.000	0.238	0.196	0.185	0.183
InternVL 3.5	0.208	0.190	0.164	0.208	0.000	0.208	0.319	0.292	0.199
Qwen3-VL-235B-A22B	0.194	0.166	0.171	0.181	0.089	0.181	0.320	0.399	0.210
GPT-5 Nano	0.143	0.140	0.137	0.143	0.000	0.143	0.502	0.478	0.211
Claude 4.5 Haiku	0.201	0.754	0.151	0.209	0.096	0.131	0.379	0.224	0.268
GPT-5 Mini	0.190	0.250	0.187	0.190	0.089	0.190	0.571	0.566	0.279
L4-Scout-17B16E	0.382	0.277	0.296	0.314	0.124	0.351	0.337	0.305	0.298
Claude Opus 4.8	0.542	0.635	0.461	0.208	0.000	0.542	0.289	0.198	0.359
Grok-4 Fast	0.357	0.564	0.281	0.312	0.222	0.350	0.461	0.344	0.361
Claude 4.5 Sonnet	0.412	0.567	0.322	0.427	0.231	0.311	0.492	0.410	0.397
Grok 4 (0709)	0.351	0.729	0.296	0.351	0.213	0.304	0.606	0.530	0.423
Gemini 2.5 Flash	0.482	0.440	0.365	0.399	0.377	0.482	0.496	0.449	0.436
Gemini 2.5 Pro	0.542	0.369	0.396	0.488	0.352	0.443	0.487	0.427	0.438
Gemini 3.1 Pro	0.673	0.603	0.597	0.350	0.132	0.626	0.296	0.360	0.455
GPT-5	0.542	0.534	0.441	0.419	0.457	0.542	0.616	0.594	0.518

Table 40: (22) Task: Hugeicons, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(23) Task: Crypto Purchase Estimation									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.021	0.051	0.018	0.019	0.000	0.016	0.262	0.118	0.063
Qwen2.5-VL-72B	0.232	0.163	0.209	0.203	0.046	0.137	0.279	0.353	0.203
Qwen3-VL-235B-A22B	0.265	0.199	0.258	0.242	0.098	0.182	0.311	0.390	0.242
Gemini 2.5 Flash Lite	0.360	0.415	0.348	0.360	0.000	0.360	0.316	0.259	0.302
Claude 4.5 Haiku	0.314	1.000	0.278	0.309	0.110	0.132	0.379	0.224	0.352
L4-Scout-17B16E	0.596	0.394	0.544	0.465	0.142	0.355	0.337	0.305	0.392
InternVL 3.5	0.524	0.654	0.480	0.524	0.251	0.185	0.319	0.292	0.404
GPT-5 Nano	0.571	0.223	0.566	0.402	0.090	0.478	0.502	0.478	0.414
Grok-4 Fast	0.557	0.803	0.517	0.461	0.254	0.353	0.461	0.344	0.469
Gemini 2.5 Pro	0.569	0.649	0.542	0.529	0.123	0.477	0.487	0.427	0.476
Claude 4.5 Sonnet	0.642	0.808	0.593	0.632	0.265	0.314	0.492	0.410	0.519
Gemini 2.5 Flash	0.635	0.696	0.590	0.479	0.245	0.577	0.496	0.449	0.521
GPT-5 Mini	0.706	0.449	0.699	0.706	0.288	0.292	0.571	0.566	0.535
GPT-5.5	0.804	0.813	0.762	0.312	0.341	0.804	0.294	0.388	0.565
Grok 4 (0709)	0.754	0.811	0.695	0.558	0.440	0.350	0.606	0.530	0.593
GPT-5	0.770	0.716	0.730	0.690	0.337	0.486	0.616	0.594	0.617
Claude Opus 4.8	0.870	0.940	0.799	0.402	0.615	0.764	0.472	0.570	0.679
Gemini 3.1 Pro	0.918	0.729	0.856	0.643	0.736	0.868	0.481	0.704	0.742

Table 41: (23) Task: Crypto Purchase Estimation, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(24) Task: Stock Growth Estimation									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.050	0.041	0.048	0.050	0.000	0.050	0.128	0.022	0.049
Gemini 2.5 Flash Lite	0.388	0.275	0.384	0.388	0.071	0.263	0.184	0.165	0.265
L4-Scout-17B16E	0.575	0.156	0.562	0.525	0.198	0.187	0.287	0.412	0.363
Qwen2.5-VL-72B	0.531	0.444	0.514	0.531	0.306	0.206	0.234	0.312	0.385
Qwen3-VL-235B-A22B	0.544	0.468	0.538	0.559	0.341	0.245	0.271	0.357	0.414
Claude 4.5 Haiku	0.563	0.361	0.552	0.563	0.243	0.365	0.355	0.331	0.417
Gemini 2.5 Flash	0.613	0.436	0.604	0.600	0.329	0.613	0.390	0.525	0.514
InternVL 3.5	0.713	0.479	0.704	0.688	0.391	0.604	0.355	0.419	0.544
Grok-4 Fast	0.688	0.359	0.680	0.688	0.519	0.658	0.380	0.479	0.556
GPT-5 Mini	0.925	0.474	0.911	0.838	0.682	0.603	0.350	0.717	0.687
GPT-5 Nano	0.950	0.361	0.937	0.900	0.813	0.752	0.358	0.580	0.706
Claude 4.5 Sonnet	0.950	0.748	0.934	0.925	0.859	0.446	0.372	0.502	0.717
Gemini 2.5 Pro	0.838	0.729	0.811	0.813	0.662	0.807	0.507	0.576	0.718
GPT-5	0.888	0.375	0.885	0.875	0.695	0.793	0.512	0.725	0.718
Claude Opus 4.8	0.900	0.778	0.873	0.662	0.708	0.832	0.369	0.670	0.724
Grok 4 (0709)	0.875	0.585	0.871	0.863	0.671	0.832	0.505	0.675	0.735
Gemini 3.1 Pro	0.912	0.685	0.890	0.863	0.760	0.820	0.385	0.700	0.752
GPT-5.5	1.000	0.830	0.975	0.925	1.000	1.000	0.495	0.912	0.892

Table 42: (24) Task: Stock Growth Estimation, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(25) Task: FIPE Car Variant Lookup									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.033	0.028	0.030	0.033	0.000	0.033	0.204	0.126	0.061
Gemini 2.5 Flash Lite	0.033	0.167	0.033	0.033	0.000	0.033	0.258	0.089	0.081
Claude 4.5 Haiku	0.200	0.184	0.191	0.200	0.024	0.133	0.355	0.244	0.191
InternVL 3.5	0.200	0.299	0.188	0.200	0.048	0.134	0.367	0.255	0.211
L4-Scout-17B16E	0.333	0.086	0.316	0.333	0.143	0.062	0.267	0.320	0.233
Grok-4 Fast	0.333	0.506	0.312	0.283	0.048	0.239	0.442	0.386	0.319
Gemini 2.5 Flash	0.367	0.357	0.357	0.283	0.148	0.244	0.496	0.372	0.328
Qwen2.5-VL-72B	0.500	0.349	0.474	0.367	0.191	0.435	0.346	0.347	0.376
Claude 4.5 Sonnet	0.500	0.397	0.472	0.450	0.167	0.162	0.492	0.467	0.388
GPT-5 Nano	0.567	0.229	0.535	0.467	0.191	0.338	0.471	0.327	0.391
GPT-5 Mini	0.467	0.305	0.445	0.467	0.167	0.307	0.513	0.550	0.403
Grok 4 (0709)	0.533	0.295	0.516	0.433	0.238	0.249	0.508	0.486	0.407
Qwen3-VL-235B-A22B	0.519	0.382	0.505	0.399	0.231	0.465	0.379	0.386	0.407
Gemini 2.5 Pro	0.500	0.570	0.483	0.450	0.119	0.248	0.646	0.534	0.444
GPT-5	0.700	0.216	0.640	0.450	0.071	0.398	0.600	0.500	0.447
GPT-5.5	0.967	0.320	0.934	0.917	0.833	0.596	0.567	0.917	0.756
Claude Opus 4.8	0.967	0.775	0.924	0.850	0.849	0.757	0.733	0.861	0.840
Gemini 3.1 Pro	1.000	0.652	0.963	0.850	1.000	0.608	0.679	0.973	0.841

Table 43: (25) Task: FIPE Car Variant Lookup, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(26) Task: Class Demand Estimation									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.117	0.052	0.117	0.117	0.000	0.117	0.354	0.294	0.146
Claude 4.5 Haiku	0.167	0.166	0.160	0.167	0.030	0.167	0.321	0.197	0.172
InternVL 3.5	0.150	0.233	0.150	0.150	0.000	0.150	0.350	0.349	0.192
Gemini 2.5 Flash Lite	0.222	0.111	0.217	0.160	0.000	0.222	0.279	0.333	0.193
Claude 4.5 Sonnet	0.278	0.384	0.274	0.215	0.000	0.278	0.429	0.297	0.269
Qwen2.5-VL-72B	0.442	0.191	0.439	0.442	0.015	0.386	0.396	0.436	0.343
L4-Scout-17B16E	0.442	0.226	0.439	0.442	0.141	0.442	0.371	0.400	0.363
Qwen3-VL-235B-A22B	0.463	0.230	0.469	0.465	0.067	0.414	0.425	0.470	0.379
GPT-5 Nano	0.397	0.622	0.392	0.397	0.181	0.206	0.438	0.501	0.392
Gemini 2.5 Flash	0.464	0.544	0.460	0.436	0.000	0.464	0.467	0.416	0.406
Gemini 2.5 Pro	0.558	0.418	0.547	0.376	0.061	0.558	0.521	0.502	0.443
Grok-4 Fast	0.544	0.722	0.544	0.482	0.000	0.544	0.517	0.475	0.478
GPT-5 Mini	0.540	0.325	0.526	0.518	0.366	0.447	0.567	0.568	0.482
GPT-5.5	0.667	0.531	0.651	0.639	0.449	0.504	0.375	0.528	0.543
Grok 4 (0709)	0.614	0.528	0.610	0.586	0.167	0.614	0.625	0.707	0.556
GPT-5	0.856	0.456	0.799	0.681	0.606	0.387	0.653	0.598	0.630
Claude Opus 4.8	0.806	0.713	0.703	0.562	0.677	0.806	0.754	0.563	0.698
Gemini 3.1 Pro	0.833	0.581	0.731	0.833	0.722	0.733	0.621	0.680	0.717

Table 44: (26) Task: Class Demand Estimation, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

(27) Task: NixOS Diagnostic									
Model	Detection		Trajectory Alignment				LLM Judge		Avg Score
	Recall	Prec	Arg Sim	Step Coh	Ord Cons	Merge Pur	Task Comp	Info Grnd	
GLM 4.5v	0.000	0.000	0.000	0.000	0.000	0.000	0.225	0.042	0.033
Claude Opus 4.8	0.069	0.059	0.060	0.049	0.000	0.069	0.375	0.500	0.148
L4-Scout-17B16E	0.222	0.121	0.213	0.188	0.045	0.173	0.221	0.160	0.168
InternVL 3.5	0.167	0.413	0.157	0.132	0.000	0.167	0.279	0.083	0.175
GPT-5.5	0.167	0.185	0.137	0.125	0.024	0.167	0.296	0.485	0.198
Gemini 2.5 Flash Lite	0.153	0.536	0.132	0.118	0.024	0.112	0.254	0.301	0.204
Gemini 3.1 Pro	0.194	0.274	0.185	0.174	0.000	0.194	0.246	0.367	0.204
Claude 4.5 Haiku	0.306	0.329	0.294	0.306	0.053	0.252	0.383	0.354	0.285
GPT-5 Nano	0.306	0.821	0.301	0.292	0.049	0.215	0.312	0.354	0.331
Grok-4 Fast	0.319	0.571	0.316	0.271	0.069	0.319	0.354	0.496	0.339
Qwen2.5-VL-72B	0.389	0.647	0.378	0.354	0.041	0.389	0.271	0.361	0.354
Claude 4.5 Sonnet	0.403	0.585	0.354	0.375	0.150	0.219	0.362	0.409	0.357
Grok 4 (0709)	0.292	0.964	0.274	0.292	0.163	0.292	0.444	0.416	0.392
Qwen3-VL-235B-A22B	0.421	0.652	0.418	0.386	0.094	0.414	0.313	0.394	0.393
Gemini 2.5 Flash	0.333	0.804	0.295	0.333	0.155	0.333	0.400	0.548	0.400
GPT-5 Mini	0.597	0.538	0.590	0.542	0.191	0.356	0.446	0.549	0.476
Gemini 2.5 Pro	0.569	0.538	0.512	0.399	0.268	0.472	0.533	0.524	0.477
GPT-5	0.764	0.392	0.747	0.572	0.499	0.556	0.421	0.583	0.567

Table 45: (27) Task: NixOS Diagnostic, Cross-Model Evaluation of Multimodal MCP Tool-Calling Trajectories.

G RANKING STABILITY

We assess ranking stability with 5,000 trajectory-level bootstrap resamples of the 211 evaluation cases for the 14 models in the original leaderboard. For each of the six alignment-based metrics, we compare resampled model rankings with the full-data ranking using Kendall's τ and report how often the full-data top model remains first.

Table 46: Ranking stability over the original 14 models under 5,000 trajectory-level bootstrap resamples.

Metric	Kendall's τ	Top-1 retained
Recall	0.910	98.6%
Precision	0.875	90.2%
Argument Similarity	0.845	100.0%
Step Coherence	0.887	91.0%
Merge Purity	0.890	81.8%
Order Consistency	0.877	95.3%
Mean	0.881	92.8%