HAZARDWEAVER: SCIENTIFIC ROUTE SELECTION FOR HAZARD ANALYSIS AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and assessing natural hazards is essential for disaster preparedness and risk reduction. Recent advances in large language models have spurred growing interest in AI agents for hazard analysis, particularly their ability to integrate scientific data, models, and tools into automated workflows. However, effective automation requires agents to determine which scientific methods are appropriate for a given event and executable with the available data and tools. As new evidence and execution results become available, these conditions can change, requiring agents to reconsider their choices. We formulate this problem as state-dependent scientific route selection and introduce HazardWeaver. Specifically, HazardWeaver first leverages the Hazard Knowledge Compiler to extract evidence-linked conditions governing scientific applicability, then its Hazard Capability Graph represents executable scientific capabilities and checks compatibility between their inputs and outputs. Using these complementary representations, the Hazard Weaver Agent component selects applicable and executable routes, carries out their workflows, and revises its decisions as the analysis state changes. To evaluate both the scientific outputs and the decisions that produce them, we introduce the Hazard Weaver Benchmark, comprising 141 instances across seven single-hazard domains and four multi-hazard interaction classes. The benchmark accommodates multiple valid scientific routes and evaluates output correctness, route validity, and justified abstention. Extensive experiments on this benchmark show that HazardWeaver outperforms existing agent systems, with the largest gains on tasks with multiple eligible scientific routes. Our code is publicly available at `https://github.com/HazardWeaver2027/Hazard_Weaver`.

1 INTRODUCTION

Natural-hazard analysis supports disaster preparedness and risk reduction. The same target can often be reached through several scientific routes. Wildfire spread, flood extent, and earthquake source parameters, for example, can each be estimated through empirical, physical, statistical, or learned methods, which differ in assumptions, required observations, validity ranges, and computational dependencies (Perry, 1998; Huot et al., 2022; Nearing et al., 2024; Zhu & Beroza, 2018; Stewart et al., 2015; Reichstein et al., 2019; Karpatne et al., 2017). A task specification alone therefore does not fix how it should be analyzed, since scientific evidence determines which routes apply while the available resources determine which routes can currently run. The difficulty grows when multiple hazards occur simultaneously or in cascade, since one hazard alters the conditions under which another must be assessed (Kappes et al., 2012; Gill & Malamud, 2014; Zscheischler et al., 2018; 2020). Recent advances in large language models have spurred growing interest in AI agents that integrate such scientific data, models, and tools into automated workflows (Boiko et al., 2023; Bran et al., 2024). On an actual event, however, the evidence and tools an agent depends on can shift as the event unfolds. The agent must therefore repeatedly reconsider which route it can trust and run. Route choice thus becomes a state-dependent decision, which leads to the question we study: given a hazard task and the current analysis state, which scientific route should an agent take?

Existing work provides important components for automated hazard analysis, but two gaps remain in how scientific routes are grounded and revised. *Gap 1: Scientific Route Grounding.* Hazard models offer established procedures within selected modeling regimes (Finney, 1998; Bates et al., 2010), yet the conditions governing their applicability are scattered across papers, reports, and model

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.

documentation. Selecting an appropriate route therefore requires making these conditions explicit and checking them against the available task evidence. This grounding must also extend to how scientific assets are combined. Although workflow and disaster-agent systems expose executable tools and operational pipelines, a route may draw on heterogeneous assets whose outputs and inputs are not directly compatible. These assets must therefore be represented as executable capabilities with explicit applicability conditions and compatibility requirements. *Gap 2: State-Dependent Route Selection.* Automated experimentation methods, from classical AutoML search (Akiba et al., 2019; Wang et al., 2021) to recent LLM-driven data-science agents, optimize models, programs, or workflows for a specified objective. However, this optimization alone does not establish which routes remain scientifically and operationally eligible as execution proceeds. In hazard analysis, new artifacts, unavailable inputs, execution failures, or changing downstream requirements can alter the eligible route set. An agent must therefore reassess route eligibility as the scientific and computational state evolves and revise its decisions accordingly. These requirements also call for route-aware evaluation: because several routes may validly satisfy the same task, evaluation must assess both final-artifact quality and whether route selection, revision, or abstention was justified under the observed state.

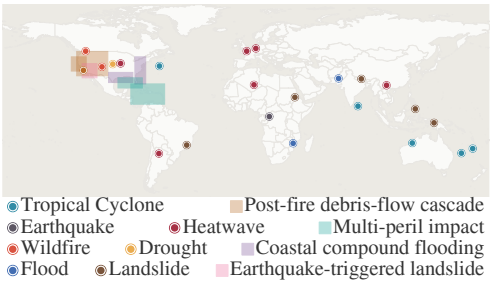


Figure 1: Global coverage of HWB.

To meet these requirements, HAZARDWEAVER integrates scientific grounding, state-dependent route selection, and route-aware evaluation within a unified evidence-to-execution architecture. Scientific route grounding is supported by two complementary components. The Hazard Knowledge Compiler (HKC) derives evidence-linked applicability conditions from scientific literature and resolves them against the task context. The Hazard Capability Graph (HCG) represents hazard models, datasets, formulas, and transformations as typed executable capabilities, connecting them only under a registered compatibility relation. Building on these components, the Hazard Weaver Agent (HWA) determines the eligible route set by combining the applicability conditions from HKC with capability reachability in HCG. It selects and executes a route, then reassesses eligibility and revises its decisions as the state evolves. Within this process, the language model manages route decisions and execution flow, while scientific models and solvers produce the numerical, spatial, and probabilistic outputs. To evaluate both the resulting artifacts and the decisions that produce them, the Hazard Weaver Benchmark (HWB) assesses artifact correctness, route validity at commitment time, and recovery-aware abstention. Specifically, HWB contains 141 sealed instances spanning seven single-hazard domains and four multi-hazard interaction classes. Its construction infrastructure includes a global atlas of 3,062 registered geospatial scenarios, with 542 materialized for benchmark generation (Figure 1).

Contributions. Our contributions can be summarized as follows:

- **A novel scientific-agent decision problem.** We formulate hazard analysis as state-dependent scientific route selection, where route eligibility requires both evidence-based scientific applicability and capability reachability, and is updated throughout execution.
- **An evidence-grounded adaptive agent system.** We introduce an agent system that compiles evidence into route conditions, connects scientific capabilities through typed compatibility, and updates executable workflows as the analysis state evolves.
- **A globally grounded, route-aware benchmark.** HWB comprises 141 instances built from scientific datasets and benchmarks, spans seven single-hazard domains and four multi-hazard interaction classes, and jointly evaluates artifact correctness and route validity.

2 PRELIMINARIES

Let $\mathcal{H}$ denote the set of hazard settings, each being either a single hazard or a multi-hazard interaction class. For $h \in \mathcal{H}$, let $\mathcal{T}_h$ denote its associated scientific tasks and $\mathcal{R}_t$ the registered route families for task $t \in \mathcal{T}_h$, together forming the design space $\mathcal{M} = \{(h, t, r) : h \in \mathcal{H},\ t \in \mathcal{T}_h,\ r \in \mathcal{R}_t\}$. A task instance is $q = (h, t, \xi, \Gamma_t, c_q)$, where ξ is the event and environmental context, Γ_t specifies

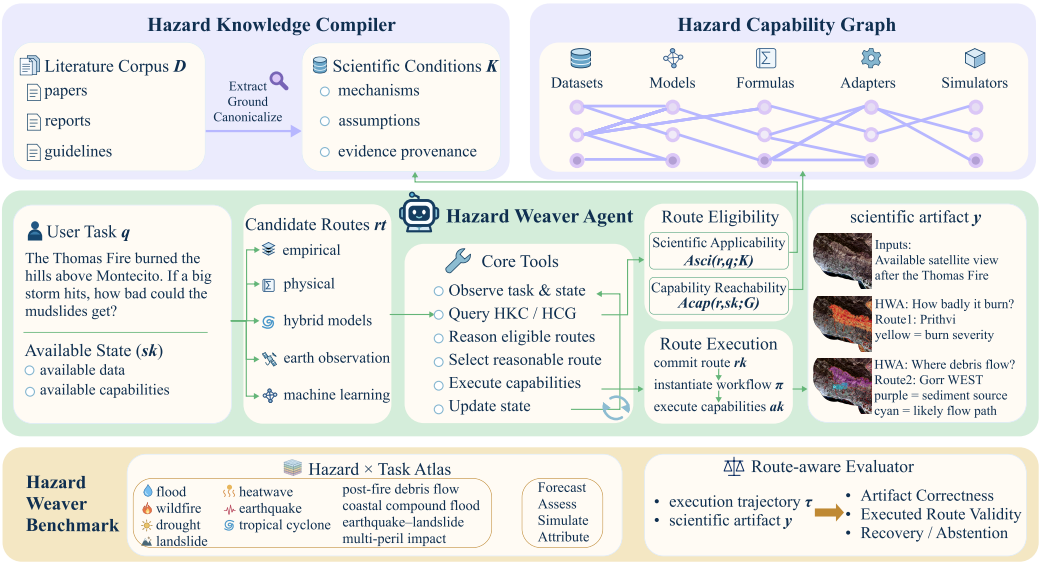


Figure 2: Overview of HazardWeaver. HKC and HCG identify scientifically applicable and currently executable routes, whose intersection defines the eligible route set for HWA. HWA selects, executes, and revises model-backed routes as the analysis state changes, while HWB independently evaluates the resulting artifact, route validity, and recovery or abstention.

the requested output in $\mathcal{Y}_t$, and c_q is the information cutoff associated with the decision. We denote scientific knowledge by $\mathcal{K}$ and executable scientific capabilities by $\mathcal{G}$.

Definition 1 (Route family and workflow)**.** A route family $r \in \mathcal{R}_t$ specifies a scientific solution structure together with its scientific conditions and required capability signatures. A concrete workflow $\pi \in \mathrm{Inst}(r; \mathcal{G})$ realizes r by binding its stages to registered datasets, models, formulas, transformations, software, and other capabilities in $\mathcal{G}$.

Definition 2 (Analysis state and trajectory)**.** At decision step k, the analysis state is $s_k = (\mathcal{X}_k, \mathcal{C}_k, b_k)$, where $\mathcal{X}_k$ contains available raw and derived artifacts, $\mathcal{C}_k$ records capability availability, and b_k is the remaining resource budget. An execution trajectory is $\tau = (s_0, a_0, s_1, \ldots, a_{K-1}, s_K)$, and route commitments made along τ are recorded as $\mathrm{Commit}(\tau) = \{(k, r_k)\}$.

A route counts as usable at s_k only when it clears two checks: scientific applicability $A_{\mathrm{sci}}(r, q, s_k; \mathcal{K})$, resolved from the evidence attached to r, and capability reachability $A_{\mathrm{cap}}(r, s_k; \mathcal{G})$, resolved from the executable assets available at s_k; Section 3 details how each operator is computed. The eligible route set at s_k collects exactly the routes passing both checks,

$$\mathcal{R}_t^{\mathrm{elig}}(q, s_k) \;=\; \Big\{ r \in \mathcal{R}_t \;:\; A_{\mathrm{sci}}(r, q, s_k; \mathcal{K}) = \textsc{Applicable} \;\wedge\; A_{\mathrm{cap}}(r, s_k; \mathcal{G}) = \textsc{Reachable} \Big\}. \tag{1}$$

Problem 1 (State-Dependent Scientific Route Selection and Execution)**.** Given $(q, s_0, \mathcal{K}, \mathcal{G}, B)$, the goal at every committed step k is to execute a route r_k drawn from $\mathcal{R}_t^{\mathrm{elig}}(q, s_k)$ (Eq. 1), recomputing this set as s_k evolves with new evidence, artifacts, and capability availability, and returning $y \in \mathcal{Y}_t$ or abstaining with $\perp$ once no state reachable within B still admits one.

3 Methodology

This section develops HazardWeaver, shown in Figure 2, in the order its route decision actually forms. Section 3.1 grounds scientific applicability through HKC, Section 3.2 establishes capability reachability through HCG, Section 3.3 combines both signals in HWA to select, execute, and revise a route, and Section 3.4 evaluates the resulting decisions through HWB.

3.1 HAZARD KNOWLEDGE COMPILER: SCIENTIFIC APPLICABILITY

A scientific route applies only when its assumptions and validity conditions match the task context, evidence that sits scattered across papers and model documentation. HKC extracts a structured condition set from its current literature corpus and checks it against the available task evidence. Retrieval-augmented methods provide relevant context (Lewis et al., 2020), while scientific information extraction provides representations of entities and relations (Luan et al., 2018; Jain et al., 2020).

Evidence-Grounded Condition Extraction. For each route family r, HKC retrieves this evidence, maps it into a common schema, and canonicalizes equivalent entities into a condition set $\mathcal{K}_r = \{\kappa_1, \ldots, \kappa_{m_r}\}$, where each κ_i records a proposition's semantic role, applicable scope, logical direction, and supporting source, spanning route assumptions, required observations, validity ranges, scale, uncertainty, and known failure conditions. **Condition-Level Applicability.** At task time, HKC relates each required condition $\kappa_i \in \mathcal{K}_r^{\text{req}}$ to the evidence available in ξ and s_k, assigns $z_i(q, s_k) \in$ SATISFIED, VIOLATED, UNRESOLVED, and derives route-level applicability:

$$A_{\text{sci}}(r, q, s_k; \mathcal{K}) = \begin{cases} \text{APPLICABLE}, & z_i(q, s_k) = \text{SATISFIED}, \ \forall \kappa_i \in \mathcal{K}_r^{\text{req}}, \\ \text{INAPPLICABLE}, & \exists\, \kappa_i \in \mathcal{K}_r^{\text{req}} : z_i(q, s_k) = \text{VIOLATED}, \\ \text{UNDETERMINED}, & \text{otherwise.} \end{cases} \tag{2}$$

The task target and information cutoff remain fixed; condition status is recomputed from evidence available in the current state. Applicability is thus grounded in individual condition status, each carrying its own provenance and task evidence, so an unresolved or changed condition can revise the route decision without reconstructing the rest of the record; an UNDETERMINED condition keeps r outside the eligible set (Eq. 1) without discarding it, since HWA can request the missing evidence during recovery (Section 3.3) and re-resolve A_{sci} once new context arrives.

3.2 HAZARD CAPABILITY GRAPH: SCIENTIFIC CAPABILITY COMPOSITION

Scientific applicability does not guarantee executability, since an appropriate route can still require unavailable capabilities or produce artifacts that no downstream model can consume. Analysis-readiness specifications motivate explicit scientific metadata (Baumann, 2024), and semantic workflow systems support composition under component constraints (Gil et al., 2011). HCG represents the available scientific ecosystem as a typed directed graph $\mathcal{G} = (\mathcal{V}_G, \mathcal{E}_G)$ and checks whether a route's required capabilities are actually reachable in the current state.

Typed Capability Composition. Nodes in $\mathcal{G}$ are executable capabilities, each recording its entry point, scientific function, dependencies, validation metadata, provenance, and typed input and output ports of type $\alpha = (v, u, f, \Omega_x, \Omega_t, \delta, \sigma, p)$, and edges are valid artifact transfers: for a produced artifact of type α' and a required input type α, a direct connection is permitted only when

$$\alpha' \models \alpha \quad \Longleftrightarrow \quad \bigwedge_{j=1}^{8} \alpha'_j \sim_j \alpha_j. \tag{3}$$

where $\sim_j$ is the registered compatibility relation for field j; exact agreement and representation-preserving conversions satisfy this directly, while information-changing operations such as aggregation, interpolation, learned downscaling, and thresholding remain explicit graph capabilities, so their effects stay visible in the workflow's provenance. **State-Dependent Capability Reachability.** Given route family r and state s_k, HCG searches for workflow realizations $\text{Inst}(r; \mathcal{G}, s_k)$ whose required capabilities are available and whose adjacent artifacts satisfy Eq. 3, giving capability reachability

$$A_{\text{cap}}(r, s_k; \mathcal{G}) = \begin{cases} \text{REACHABLE}, & \text{Inst}(r; \mathcal{G}, s_k) \neq \varnothing, \\ \text{UNAVAILABLE}, & \text{Inst}(r; \mathcal{G}, s_k) = \varnothing \text{ under the known state}, \\ \text{UNDETERMINED}, & \text{capability information is incomplete.} \end{cases} \tag{4}$$

As with A_{sci}, an UNDETERMINED status keeps r outside the eligible set without discarding it. When capability availability or an artifact changes, HCG searches for a compatible replacement subpath or prompts HWA to reconsider the route when it does not.

Provenance records connect each derived artifact to the capabilities and inputs that produced it, following the purpose of interoperable workflow provenance (Khan et al., 2019). The illustrated scientific assets include an Earth-observation foundation model (Jakubik et al., 2023) and WEST, a postfire debris-flow volume model (Gorr et al., 2026). A spatial flow-path artifact requires a separate downstream capability in addition to a volume estimate.

3.3 HAZARD WEAVER AGENT: ADAPTIVE ROUTE SELECTION

HWA operationalizes Eq. 1 at every decision step, turning the intersection of scientific applicability and capability reachability into an executed, revisable route commitment.

Model-Backed Route Execution. At state s_k, $\mathcal{R}_t^{\mathrm{elig}}(q, s_k)$ combines A_{sci} from HKC with A_{cap} from HCG, which exclude routes for different reasons, conflicting scientific requirements versus requirements that cannot currently be realized, letting HWA trace an excluded route to its specific cause and target the matching recovery operation. For every r in this set, HWA receives its grounded conditions, workflow realizations, provenance, uncertainty, and resource requirements, and a language-model policy selects $r_k \sim \mu_\theta(\cdot \mid q, s_k, \mathcal{R}_t^{\mathrm{elig}}(q, s_k))$, records the commitment (k, r_k), instantiates $\pi^{(k)} \in \mathrm{Inst}(r_k; \mathcal{G}, s_k)$, and executes its capabilities; the language model controls only the route decision and execution flow, while the executed models, solvers, formulas, and transformations produce the numerical, spatial, and probabilistic outputs, keeping the terminal artifact linked through provenance to the route that produced it. **Revisable Route Commitments.** Execution can reveal missing inputs, alter capability availability, or invalidate downstream compatibility, so HWA updates s_k and recomputes Eq. 1 after every step: a route that remains eligible continues with its current or repaired workflow, and one that does not sends the policy back to the updated eligible set. When this set is empty, HWA explores permitted recovery operations, information requests, compatible transformations, capability substitutions, and alternative realizations, over the states $\mathrm{Rec}(s_k, b_k)$ reachable within the remaining budget, and uses the recovery outcome to decide whether to abstain. Here $\mathrm{Rec}(s_k, b_k)$ is the budget-bounded reachable closure over the finite set of grounded actions permitted for the instance. The implemented search explores this closure within its budget. Failure to find a route is not a formal proof that no feasible route exists.

3.4 HAZARD WEAVER BENCHMARK: ROUTE-AWARE EVALUATION

A task may admit several valid routes, and the valid set may shift as conditions or capabilities change, so artifact correctness alone cannot certify a valid decision. We therefore introduce the **Hazard Weaver Benchmark** (HWB), which evaluates the final artifact together with the route and execution state that produced it. HWB derives tasks with explicit output contracts from established scientific datasets, operational products, and benchmarks, and constructs candidate routes from evidence-grounded Route Cards and registered capabilities. Instances are admitted only after provenance, temporal-cutoff, and replay checks, after which the inventory, route registry, and task-specific evaluators are sealed; benchmark-side references and route judgments remain hidden from the solver. Appendix A and Appendix C detail the construction and evaluation protocol.

Benchmark-Side Route Validity. Using hidden scientific conditions and benchmark-side capability state kept outside solver context, HWB constructs $\mathcal{R}_t^{\mathrm{elig},B}(q, s_k)$ by the same criteria as Eqs. 2–4. A trajectory is route-valid when every commitment falls inside this set at its commitment state,

$$V_{\mathrm{route}}^{B}(q, \tau) = \mathbb{1}\left[r_k \in \mathcal{R}_t^{\mathrm{elig},B}(q, s_k),\ \forall (k, r_k) \in \mathrm{Commit}(\tau)\right], \tag{5}$$

crediting any route the benchmark itself registers as valid without requiring the agent to reproduce the construction-time witness workflow. Writing $\theta(x)$ for a raw artifact's information time and $\mathrm{Src}(z)$ for the raw-source ancestors of derived artifact z, temporal validity requires every such ancestor to have been available by the cutoff c_q,

$$V_{\mathrm{causal}}(q, \tau) = \mathbb{1}\Big[\theta(x) \le c_q,\ \forall x \in \textstyle\bigcup_{z \in \mathrm{Used}(\tau)} \mathrm{Src}(z)\Big], \tag{6}$$

so a derived artifact may be computed after c_q once all its sources were available in time. Together with $V_{\mathrm{type}}(\tau)$ from Eq. 3, provenance validity $V_{\mathrm{prov}}(y, \tau)$ linking the terminal artifact to an executed route commitment, and budget validity $V_{\mathrm{budget}}(\tau; B)$, these give joint execution validity $V_{\mathrm{exec}}(q, \tau; B) = V_{\mathrm{type}}(\tau) \cdot V_{\mathrm{causal}}(q, \tau) \cdot V_{\mathrm{budget}}(\tau; B)$. The benchmark contract accepts abstention only when permitted recovery cannot reach another registered benchmark-valid route,

$$V_{\mathrm{abs}}^{B}(q, s_k, b_k) = \mathbb{1}\Big[\textstyle\bigcup_{s' \in \mathrm{Rec}(s_k, b_k)} \mathcal{R}_t^{\mathrm{elig},B}(q, s') = \varnothing\Big], \tag{7}$$

which defines the benchmark criterion over its registered recovery transitions. Applying this criterion requires the benchmark to establish the reachable set under those transitions; a solver timeout alone does not establish that no usable route remains.

Decision-Constrained Accuracy. HWB combines artifact correctness with these validity checks into Decision-Constrained Accuracy: writing the solver's return as $o \in \mathcal{Y}_t \cup \{\perp\}$ and $E_q(y) \in \{0, 1\}$ for task-specific artifact success under its native evaluator,

$$\mathrm{DCA}(q, o, \tau; B) = \begin{cases} E_q(y)\, V^B_{\mathrm{route}}(q, \tau)\, V_{\mathrm{prov}}(y, \tau)\, V_{\mathrm{exec}}(q, \tau; B), & o = y, \\ V^B_{\mathrm{abs}}(q, s_K, b_K)\, V^B_{\mathrm{route}}(q, \tau)\, V_{\mathrm{exec}}(q, \tau; B), & o = \perp, \end{cases} \tag{8}$$

aggregated over an evaluation set as $\mathrm{DCA}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{DCA}(q, o_q, \tau_q; B)$; native scientific metrics stay task-specific and are reported separately, so DCA measures decision validity without folding heterogeneous scientific errors into one score. HWB further builds matched interventions that alter one factor governing route choice while holding the rest of the task fixed: evidence relevant to A_{sci}, availability or compatibility relevant to A_{cap}, the state after execution begins, or an upstream hazard variable in a registered physical interaction. These interventions test whether route selection, recovery, and the resulting artifacts respond consistently to controlled changes in the route space.

4 Experimental Setup

Benchmark. The primary comparison uses the same 141 sealed HWB instances, HCG capability catalog, and evaluator for all systems. The instances draw on eleven published datasets, operational products, and scientific benchmarks: PWFDF and USGS Ground Failure (King et al., 2025; Allstadt et al., 2022), SFINCS and FloodCastBench (Leijnse et al., 2021; Xu et al., 2025), NOAA CPC SPI (NOAA Climate Prediction Center, 2026), SeisBench (Woollam et al., 2022), ExtremeWeatherBench and NASA LHASA v2 (McGovern et al., 2026; NASA Goddard Earth Sciences Data and Information Services Center, 2025), TCBench and NOAA Storm Events (Gomez et al., 2026; NOAA National Centers for Environmental Information, 1996), and WildfireSpreadTS (Gerard et al., 2023). These sources provide observed events, forcing data, and reference outputs used to construct route-selection and execution tasks. The inventory contains 90 single-hazard and 51 multi-hazard instances; 45 admit one route and 96 admit multiple routes. Appendix A defines the tracks and evaluation strata.

Baselines. We compare HazardWeaver with adapted versions of Plan-and-Solve (Wang et al., 2023a), Self-Consistency (Wang et al., 2023b), DisasterBench-ToT (Chen et al., 2026; Yao et al., 2023), MLE-STAR (Nam et al., 2025), AutoML-Agent (Trirat et al., 2025), Reflexion (Shinn et al., 2023), AIDE (Jiang et al., 2025), DS-Agent (Guo et al., 2024), R&D-Agent (Yang et al., 2025), and ReAct (Yao et al., 2022). All methods use the same task inventory, policy backbone, route catalog, scientific replay, and evaluator. Appendix E documents the adaptations and implementation details.

Metrics. The primary metric is Decision-Constrained Accuracy (DCA; Eq. 8). We also report artifact validity E_q, aggregate trajectory validity V_q, and Macro DCA, which weights the 11 tracks equally. Binomial rates use Wilson 95% confidence intervals (Wilson, 1927).

Implementation. The five ablation backbones are OLMo-3.1-32B, Mixtral-8x22B, Gemma-4-31B-it, Llama-3.3-70B, and DeepSeek-V4.1-Flash. The corresponding model-family reports describe OLMo (Olmo et al., 2026), Mixtral (Jiang et al., 2024), Gemma (Gemma et al., 2026), Llama (Grattafiori et al., 2024), and DeepSeek (Deepseek, 2026). The evaluated variants are listed in Appendix D.

5 Experiments

We evaluate HazardWeaver through three research questions. **RQ1:** How does HazardWeaver compare with representative agent baselines? **RQ2:** Which design mechanisms contribute to decision-valid scientific execution? **RQ3:** How consistently does HazardWeaver perform across language-model backbones and hazard-task settings?

5.1 System Performance

We answer **RQ1** by comparing HazardWeaver with adapted agent systems and general control strategies. All methods use the same task inventory, Llama backbone, registered route catalog, and DCA evaluator. Table 1 reports performance across hazard composition and route multiplicity.

The comparison covers two complementary baseline families. AIDE, DS-Agent, R&D-Agent, MLE-STAR, AutoML-Agent, and DisasterBench-ToT adapt representative agent-system designs to the

Table 1: Comparison with agent baselines. Each cell reports DCA (%) with the upper and lower offsets of its Wilson 95% confidence interval. Methods are ordered by overall DCA.

Method	Single-hazard	Multi-hazard	One route	Multiple routes	Overall DCA
AIDE	$22.2^{+9.7}_{-7.3}$	$23.5^{+13.3}_{-9.5}$	$28.9^{+14.5}_{-11.2}$	$19.8^{+9.1}_{-6.7}$	$22.7^{+7.6}_{-6.1}$
AutoML-Agent	$26.7^{+9.9}_{-8.1}$	$27.4^{+13.6}_{-10.3}$	$24.4^{+14.3}_{-10.2}$	$28.1^{+9.7}_{-8.0}$	$26.9^{+7.9}_{-6.6}$
ReAct	$28.9^{+10.1}_{-8.4}$	$31.4^{+13.6}_{-11.1}$	$31.1^{+14.6}_{-11.6}$	$29.2^{+9.7}_{-8.2}$	$29.8^{+8.0}_{-6.9}$
Reflexion	$30.0^{+10.1}_{-8.5}$	$35.3^{+13.7}_{-11.7}$	$51.1^{+13.9}_{-14.1}$	$22.9^{+9.4}_{-7.2}$	$31.9^{+8.1}_{-7.1}$
MLE-STAR	$35.6^{+10.3}_{-9.2}$	$29.4^{+13.6}_{-10.7}$	$35.6^{+14.6}_{-12.4}$	$32.3^{+9.9}_{-8.5}$	$33.3^{+8.2}_{-7.2}$
DS-Agent	$33.3^{+10.3}_{-8.9}$	$33.3^{+13.7}_{-11.3}$	$40.0^{+14.5}_{-13.0}$	$30.2^{+9.8}_{-8.3}$	$33.3^{+8.2}_{-7.2}$
R&D-Agent	$33.3^{+10.3}_{-8.9}$	$35.3^{+13.7}_{-11.7}$	$42.2^{+14.5}_{-13.2}$	$30.2^{+9.8}_{-8.3}$	$34.0^{+8.2}_{-7.3}$
Self-Consistency	$33.3^{+10.3}_{-8.9}$	$39.2^{+13.7}_{-12.2}$	$42.2^{+14.5}_{-13.2}$	$32.3^{+9.9}_{-8.5}$	$35.5^{+8.1}_{-7.4}$
DisasterBench-ToT	$35.6^{+10.3}_{-9.2}$	$35.3^{+13.7}_{-11.7}$	$42.2^{+14.5}_{-13.2}$	$32.3^{+9.9}_{-8.5}$	$35.5^{+8.1}_{-7.4}$
Plan-and-Solve	$41.1^{+10.3}_{-9.6}$	$41.2^{+13.6}_{-12.4}$	$57.8^{+13.2}_{-14.5}$	$33.3^{+9.9}_{-8.6}$	$41.1^{+8.3}_{-7.7}$
HazardWeaver	$\mathbf{91.1}^{+4.3}_{-7.7}$	$\mathbf{86.3}^{+6.9}_{-12.0}$	$\mathbf{80.0}^{+9.1}_{-13.8}$	$\mathbf{93.8}^{+3.3}_{-6.8}$	$\mathbf{89.4}^{+4.1}_{-6.2}$

benchmark interface. ReAct, Reflexion, Self-Consistency, and Plan-and-Solve evaluate alternative control strategies with the same backbone and scientific tools as HazardWeaver. HazardWeaver reaches 89.4% overall DCA, compared with 41.1% for the strongest baseline. The difference is most pronounced when several routes are available: Plan-and-Solve performs well with one route but declines sharply under multiple-route choice, while HazardWeaver rises to 93.8%. Iterative refinement, retrieval, search, reflection, and explicit planning therefore provide limited benefit when the controller must identify which scientific workflow is valid. HazardWeaver remains strong across both hazard compositions and route regimes, indicating that its advantage comes from structured route eligibility rather than a particular task subset.

5.2 Component and Mechanism Analysis

RQ2 examines whether HazardWeaver benefits from explicitly representing scientific eligibility and maintaining it as execution proceeds. We test these two design choices separately. The first experiment removes the evidence-linked conditions supplied by HKC or the typed compatibility constraints supplied by HCG. The second disables route updating on the subset of tasks for which eligibility can change during execution.

Scientific constraints. Figure 3 shows the same directional result across all five backbones: Full HazardWeaver outperforms both ablated variants. The effect is strongest for policies that can otherwise execute the selected workflow reliably. On Llama, full DCA reaches 93.8%, while removing HKC conditions or HCG typing reduces it to 45.8% and 47.9%. DeepSeek exhibits the same pattern at a smaller scale. The weaker policies retain less performance overall, but neither ablation improves any backbone. These results distinguish scientific route selection from ordinary tool use. HKC removes routes whose assumptions do not match the task context, while HCG prevents compositions whose inputs and outputs are incompatible. Without either constraint, access to the correct scientific capabilities is insufficient for selecting a valid workflow.

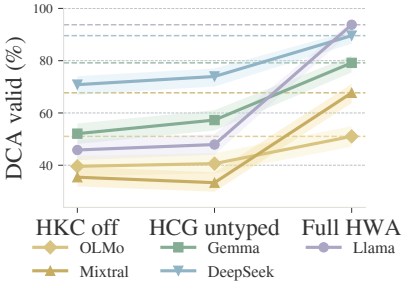


Figure 3: Scientific-constraint ablation results. DCA (%) on the 96 multiple-route instances. Shaded bands indicate Wilson 95% interval.

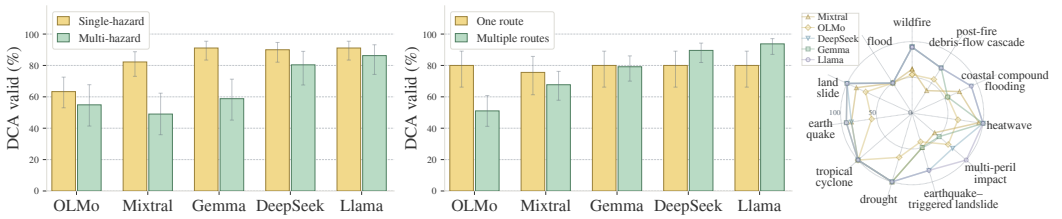
Table 2: Route update ablation. DCA (%) on 40 applicable instances.

Backbone	Fixed	Updated
Mixtral	$5.0^{+11.5}_{-3.6}$	$15.0^{+14.1}_{-7.9}$
OLMo	$17.5^{+14.4}_{-8.7}$	$25.0^{+15.2}_{-10.8}$
DeepSeek	$22.5^{+15.0}_{-10.2}$	$35.0^{+15.5}_{-12.9}$
Gemma	$20.0^{+14.8}_{-9.5}$	$42.5^{+15.3}_{-14.0}$
Llama	$15.0^{+14.1}_{-7.9}$	$90.0^{+6.0}_{-13.0}$
Pooled	16.0	**41.5**

Route updating. Table 2 isolates the second design choice on the 40 benchmark instances where route eligibility may change during execution. Recomputing eligibility improves every backbone, raising pooled DCA from 16.0% to 41.5%. The gain is largest for Llama, whose stronger policy can exploit the revised route set once an earlier workflow becomes ineligible. The consistent direction across all five backbones shows that this benefit does not depend on one model family.

Table 3: Cross-backbone performance of Full HazardWeaver. DCA is the headline metric, and Macro DCA assigns equal weight to the 11 tracks. E_q and V_q report the corresponding marginal artifact- and trajectory-validity rates under the same sealed evaluator.

Backbone	DCA	Macro DCA	Artifact validity (E_q)	Trajectory validity (V_q)
OLMo	60.3 [52.0, 68.0]	61.9 [54.4, 71.2]	78.0 [70.5, 84.1]	91.5 [85.7, 95.1]
Mixtral	70.2 [62.2, 77.1]	70.8 [57.6, 83.9]	79.4 [72.0, 85.3]	92.2 [86.6, 95.6]
Gemma	79.4 [72.0, 85.3]	78.6 [65.2, 91.5]	86.5 [79.9, 91.2]	92.9 [87.4, 96.1]
DeepSeek	86.5 [79.9, 91.2]	86.6 [77.2, 94.3]	90.8 [84.9, 94.5]	92.2 [86.6, 95.6]
Llama	**89.4** [83.2, 93.5]	**89.5** [79.7, 97.0]	**92.2** [86.6, 95.6]	**95.0** [90.1, 97.6]



(a) Hazard composition. (b) Route multiplicity. (c) Track-level coverage.

Figure 4: Cross-backbone behavior of Full HazardWeaver. Panels report DCA by (a) hazard composition, (b) route multiplicity, and (c) benchmark track.

A capable controller still requires an up-to-date decision space, since a fixed eligible set retains invalidated routes. Overall, reliable scientific route decisions require evidence-linked applicability conditions, typed capability composition, and continued maintenance of the eligible route set.

5.3 Generalization Across Backbones and Task Structures

RQ3 examines whether HazardWeaver remains effective when its policy model changes and which task structures expose the remaining differences. We hold the benchmark, HKC conditions, HCG capabilities, execution interface, and evaluator fixed, and replace only the language-model backbone.

Cross-backbone effectiveness. Table 3 shows that HazardWeaver remains effective across all five policy backbones, with overall DCA ranging from 60.3% to 89.4%. Llama and DeepSeek achieve the strongest results, while Gemma retains nearly 80% DCA with a smaller model. Macro DCA closely follows the overall ordering, showing that the ranking is balanced by all tracks. The marginal validity rates sharpen this result: artifact validity broadly follows DCA, whereas trajectory validity remains consistently high across backbones. The remaining cross-backbone separation is therefore concentrated in producing the required scientific artifact and selecting the valid scientific route.

Hazard composition. Figure 4a shows that the strongest policies preserve most of their performance when tasks move from single-hazard analysis to multi-hazard composition. Llama and DeepSeek remain above 80% on multi-hazard tasks, whereas the gap widens for Mixtral and Gemma. Scientific composition therefore provides a more discriminating test of the policy backbone than isolated hazard analysis, since the agent must carry intermediate state across stages and select a compatible downstream workflow while preserving scientific validity throughout the composed analysis.

Route multiplicity. Figure 4b reveals a sharper difference across route regimes. Performance on the one-route subset is tightly grouped across the five backbones, whereas the multiple-route subset produces substantially wider separation. Llama and DeepSeek achieve higher DCA on the multiple-route subset, while the weaker policies exploit the available alternatives less consistently. Route-rich tasks therefore provide the clearest distinction between policy backbones within the same HazardWeaver architecture. This descriptive pattern complements Section 5.2, where controlled ablations establish the contribution of the route-eligibility mechanisms themselves.

Track-level coverage. Figure 4c confirms that the aggregate results reflect broad coverage rather than success on one hazard family. The strongest backbones form consistently broad profiles across single- and multi-hazard tracks. Flood remains the clearest shared bottleneck, while compound tracks such as multi-peril impact and earthquake-triggered landslide separate the policy backbones more

strongly. These differences identify where future improvements should focus: transferring scientific state across stages and choosing among competing composed workflows.

Overall, RQ3 shows that HazardWeaver's route-selection architecture transfers across five materially different language-model backbones. Stronger policies extract greater benefit from the same scientific structure on multi-hazard and multiple-route tasks, while the shared framework preserves substantial decision-valid performance even with the weaker backbones.

6 Related Work

Scientific Route Grounding. Hazard science provides specialized empirical, physical, statistical, and learned models for single- and multi-hazard analysis, such as fire-area simulators and inertial shallow-water flood solvers (Finney, 1998; Bates et al., 2010). These models encode scientific assumptions through equations, required observations, calibration regimes, and validity ranges, while scientific workflow research represents executable operations through interfaces, dependencies, and capability graphs (Ahn et al., 2022; Zhai et al., 2026). DisasterBench evaluates planning under typed tool interfaces (Chen et al., 2026), and DORA extends disaster agents to heterogeneous geospatial and learned tools with expert-authored tasks and replayable trajectories (Wang et al., 2026). These systems establish capable execution environments, but selecting among methods that target the same output also depends on variable definitions, units, spatial and temporal support, resolution, uncertainty, provenance, and operating scale (Baumann, 2024; Lee et al., 2025). Multi-hazard analysis compounds this requirement, since the output of one hazard model becomes an input condition of the next (Kappes et al., 2012). Existing workflow representations capture executable dependencies well, while task-conditioned scientific applicability remains largely external to their route structures. HazardWeaver connects these layers by grounding alternative route families in evidence-linked conditions and composing them through typed executable capabilities.

State-Dependent Route Decision. Automated machine learning formulates estimator and pipeline selection as search over a fixed target and evaluation objective (Akiba et al., 2019; Wang et al., 2021). Recent agents extend this process to executable code, iterative experiment refinement, and data-science workflows (Shinn et al., 2023; Trirat et al., 2025; Nam et al., 2025; Huang et al., 2024), while richer planning systems organize available tools before execution (Liu et al., 2023; Kim et al., 2024; Rana et al., 2023). Hazard analysis additionally requires route eligibility to follow the current scientific state because new evidence, intermediate artifacts, and capability outcomes such as tool failures or missing inputs can change which workflows remain applicable. Related benchmarks evaluate planning, multi-step trajectories, and task outcomes (Valmeekam et al., 2023; Trivedi et al., 2024; Lu et al., 2025; Pothuri et al., 2026; Yao et al., 2024; Chan et al., 2024), including compositional and changing environments (Wang et al., 2025; Yoa et al., 2026; Zeng et al., 2025), and recent abstention benchmarks study whether agents can recognize when to stop acting (Liu et al., 2026). Route-plural scientific analysis further requires relating an executed workflow to the state in which it was selected while accommodating multiple registered solutions, and judging abstention by whether any usable route remains reachable. HazardWeaver addresses this problem by recomputing eligible scientific routes after every executed step and evaluating the resulting artifact, the validity of the committed route, and the justification of any abstention.

7 Conclusion

We introduced state-dependent scientific route selection as a central decision problem for hazard-analysis agents and presented HazardWeaver, which grounds applicability in literature evidence, composes scientific models and tools through typed capability relations, and updates eligible workflows as the analysis state evolves. The Hazard Weaver Benchmark evaluates this process on 141 sealed instances across seven single-hazard domains and four multi-hazard interaction classes. HazardWeaver achieves 89.4% Decision-Constrained Accuracy, compared with 41.1% for the strongest representative agent baseline, with its largest advantage on tasks offering multiple scientific routes. Ablating evidence grounding, capability typing, or route updating consistently degrades performance, while evaluation across five backbones confirms that the architecture transfers beyond a single policy model. These results establish explicit, state-aware scientific route selection as a practical foundation for reliable model-backed hazard analysis.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity and readability of the manuscript. All AI-assisted content was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work involves no human subjects or personally identifiable information and uses only public scientific data. HazardWeaver is a research prototype and has not been validated for operational forecasting or early warning. Provenance records make its outputs auditable, and abstention withholds outputs when no valid route remains, reducing the risk of confident but invalid analyses. Operational use would require local calibration, expert review, and human oversight.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental setup, datasets, evaluation protocols, and implementation details in the main paper and the appendix. Additional implementation details are available in the GitHub repository linked in the abstract.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, August 2022.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330701.

Kate E. Allstadt, Eric M. Thompson, Randall W. Jibson, David J. Wald, Mike Hearne, Edward J. Hunter, Jeremy Fee, Heather Schovanec, Daniel Slosky, and Kirstie Lafon Haynie. The US Geological Survey ground failure product: Near-real-time estimates of earthquake-triggered landslides and liquefaction. *Earthquake Spectra*, 38(1):5–36, 2022. doi: 10.1177/87552930211032685.

Paul D. Bates, Matthew S. Horritt, and Timothy J. Fewtrell. A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, 387(1-2):33–45, June 2010. ISSN 00221694. doi: 10.1016/j.jhydrol.2010.03.027.

Peter Baumann. On the analysis-readiness of spatio-temporal Earth data and suggestions for its enhancement. *Environmental Modelling & Software*, 176:106017, May 2024. ISSN 13648152. doi: 10.1016/j.envsoft.2024.106017.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Zhitong Chen, Kai Yin, Weifeng Zhang, Zhiyuan Wang, Xiangjue Dong, Chengkai Liu, Zhewei Liu, Yiming Xiao, Ali Mostafavi, and James Caverlee. DisasterBench: Benchmarking LLM Planning under Typed Tool Interface Constraints, May 2026.

Deepseek. DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression. September 2026.

Mark A. Finney. FARSITE: Fire Area Simulator-model development and evaluation. Technical Report RMRS-RP-4, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ft. Collins, CO, 1998.

Team Gemma, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu-hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I.-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak

Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 Technical Report, July 2026.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro Gonzalez-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems*, 26(1):62–72, January 2011. ISSN 1941-1294. doi: 10.1109/MIS.2010.9.

Joel C. Gill and Bruce D. Malamud. Reviewing and visualizing the interactions of natural hazards. *Reviews of Geophysics*, 52(4):680–722, 2014. ISSN 1944-9208. doi: 10.1002/2013RG000445.

Milton Gomez, Marie McGraw, Saranya Ganesh S, Frederick Iat-Hin Tam, Ilia Azizi, Samuel Darmon, Monika Feldmann, Stella Bourdin, Louis Poulain–Auzéau, Suzana J. Camargo, Jonathan Lin, Dan Chavas, Chia-Ying Lee, Ritwik Gupta, Andrea Jenney, and Tom Beucler. TCBench: A Benchmark for Tropical Cyclone Track and Intensity Forecasting at the Global Scale, June 2026.

Alexander N. Gorr, Francis K. Rengers, Katherine R. Barnhart, Matthew A. Thomas, and Jason W. Kean. An improved empirical model for predicting postfire debris-flow volume in the western United States. *Natural Hazards and Earth System Sciences*, 26(5):2111–2132, May 2026. ISSN 1561-8633. doi: 10.5194/nhess-26-2111-2026.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 Herd of Models, November 2024.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. DS-Agent: Automated Data Science by Empowering Large Language Models with Case-Based Reasoning, May 2024.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20271–20309. PMLR, July 2024.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next Day Wildfire Spread: A Machine Learning Data Set to Predict Wildfire Spreading from Remote-Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2022.3192974.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A Challenge Dataset for Document-Level Information Extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7506–7516, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.670.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence, November 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-Driven Exploration in the Space of Code, 2025.

Melanie S. Kappes, Margreth Keiler, Kirsten Von Elverfeldt, and Thomas Glade. Challenges of analyzing multi-hazard risk: A review. *Natural Hazards*, 64(2):1925–1958, November 2012. ISSN 0921-030X, 1573-0840. doi: 10.1007/s11069-012-0294-2.

Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, October 2017. ISSN 1041-4347. doi: 10.1109/TKDE.2017.2720168.

Farah Zaib Khan, Stian Soiland-Reyes, Richard O Sinnott, Andrew Lonie, Carole Goble, and Michael R Crusoe. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. *GigaScience*, 8(11):giz095, November 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz095.

Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. An LLM Compiler for Parallel Function Calling, June 2024.

Jonathan M. King, Jaime Kostelnik, Jason W. Kean, Katherine R. Barnhart, Andrew P. Graber, Francis K. Rengers, and Matthew A. Thomas. Post-wildfire debris-flow hazard assessment (PWFDF) collection, 2025. URL `https://doi.org/10.5066/P14EWYME`. CC0 1.0 Universal.

Sujong Lee, Minwoo Roh, Hyun-Woo Jo, Kim Joon, and Woo-Kyun Lee. Machine learning-based rainfall-induced landslide susceptibility model and short-term early warning assessment in South Korea. *Landslides*, 22(8):2809–2827, August 2025. ISSN 1612-510X, 1612-5118. doi: 10.1007/s10346-025-02513-y.

Tim Leijnse, Maarten van Ormondt, Kees Nederhoff, and Ap van Dongeren. Modeling compound flooding in coastal systems using a computationally efficient reduced-physics solver: Including fluvial, pluvial, tidal, wind- and wave-driven processes. *Coastal Engineering*, 163:103796, January 2021. ISSN 0378-3839. doi: 10.1016/j.coastaleng.2020.103796.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency, September 2023.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM Agents Know When Not to Act?, July 2026.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.65.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3219–3232, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1360.

Amy McGovern, Taylor Mandelbaum, Daniel Rothenberg, Nicholas Loveday, Corey Potvin, Montgomery Flora, Linus Magnusson, Eric Gilleland, and John Allen. Extreme Weather Bench: A framework and benchmark for evaluation of high-impact weather, May 2026.

Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, Jinwoo Shin, Sercan Arik, and Tomas Pfister. MLE-STAR: Machine Learning Engineering Agent via Search and Targeted Refinement. In *Advances in Neural Information Processing Systems 38*, pp. 129290–129310, San Diego, California, USA and Mexico City, Mexico, 2025. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-3827-5. doi: 10.52202/085713-3889.

NASA Goddard Earth Sciences Data and Information Services Center. Global landslide nowcast from LHASA L4 1 day 1 km × 1 km, version 2.0.0, 2025. URL https://doi.org/10.5067/8VKQDQFFOTS3. Dataset identifier: Global_Landslide_Nowcast; Level 4; daily temporal resolution; 1 km spatial resolution.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, Shlomo Shenzis, Tadele Yednkachw Tekalign, Dana Weitzner, and Yossi Matias. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, March 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07145-1.

NOAA Climate Prediction Center. Standardized precipitation index, 2026. URL https://www.cpc.ncep.noaa.gov/products/Drought/Monitoring/spi.shtml. Operational drought-monitoring product based on the CPC Unified Precipitation Analysis, 1950–present.

NOAA National Centers for Environmental Information. Storm Events Database, 1996. URL https://www.ncei.noaa.gov/access/storm-events-database/. Version 1.0; NCEI dataset identifier 3910_03; continuously updated.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, et al. Olmo 3, April 2026.

G. L.W. Perry. Current approaches to modelling the spread of wildland fire: A review. *Progress in Physical Geography: Earth and Environment*, 22(2):222–245, June 1998. ISSN 0309-1333. doi: 10.1177/030913339802200204.

Abhinav Pothuri, Zhe Jiang, Zelin Xu, and Di Yang. GISAgentBench: A Practitioner-Sourced Benchmark for Evaluating LLM Agents on GIS Tasks, August 2026.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, and Ian Reid. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning. 2023. URL https://arxiv.org/abs/2307.06135.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, October 2023.

Jonathan P. Stewart, John Douglas, Mohammad Javanbarg, Yousef Bozorgnia, Norman A. Abrahamson, David M. Boore, Kenneth W. Campbell, Elise Delavaud, Mustafa Erdik, and Peter J. Stafford. Selection of Ground Motion Prediction Equations for the Global Earthquake Model. *Earthquake Spectra*, 31(1):19–45, February 2015. ISSN 8755-2930, 1944-8201. doi: 10.1193/013013EQS017M.

Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 60099–60146. PMLR, October 2025.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change, November 2023.

Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. FLAML: A Fast and Lightweight AutoML Library. In *Proceedings of the Fourth Conference on Machine Learning and Systems (MLSys 2021)*, volume 3, pp. 434–447, 2021.

Junjue Wang, Weihao Xuan, Heli Qi, Pengyu Dai, Kunyi Liu, Hongruixuan Chen, Zhuo Zheng, Junshi Xia, Stefano Ermon, and Naoto Yokoya. Can LLM Agents Respond to Disasters? Benchmarking Heterogeneous Geospatial Reasoning in Emergency Operations, 2026.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, September 2023b.

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, August 2025.

Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, June 1927. ISSN 0162-1459. doi: 10.1080/01621459.1927.10502953.

Jack Woollam, Jannes Münchmeyer, Frederik Tilmann, Andreas Rietbrock, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, Alberto Michelini, Joachim Saul, and Hugo Soto. SeisBench—A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, 93(3):1695–1709, March 2022. ISSN 0895-0695. doi: 10.1785/0220210324.

Qingsong Xu, Yilei Shi, Jie Zhao, and Xiao Xiang Zhu. FloodCastBench: A Large-Scale Dataset and Foundation Models for Flood Modeling and Forecasting. *Scientific Data*, 12(1):431, March 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04725-2.

Xu Yang, Xiao Yang, Shikai Fang, Yifei Zhang, Jian Wang, Bowen Xian, Qizheng Li, Jingyuan Li, Minrui Xu, Yuante Li, Haoran Pan, Yuge Zhang, Weiqing Liu, Yelong Shen, Weizhu Chen, and Jiang Bian. R&D-Agent: An LLM-Agent Framework Towards Autonomous Data Science, October 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. doi: 10.52202/075280-0517.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, 2024.

Seungdong Yoa, Sanghyu Yoon, Suhee Yoon, Dongmin Kim, Ye Seul Sim, Junhyun Lee, and Woohyung Lim. From Static Benchmarks to Dynamic Protocol: Agent-Centric Text Anomaly Detection for Evaluating LLM Reasoning, February 2026.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Shuang Ni, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction, September 2025.

Zian Zhai, Xingyu Tan, Gaowang Zou, Xiaoyang Wang, and Wenjie Zhang. HyperAgent: Planning and Acting over Tool-Schema Hypergraphs for Tool-Use LLM Agents, July 2026.

Weiqiang Zhu and Gregory C Beroza. PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. *Geophysical Journal International*, October 2018. ISSN 0956-540X, 1365-246X. doi: 10.1093/gji/ggy423.

Jakob Zscheischler, Seth Westra, Bart J. J. M. van den Hurk, Sonia I. Seneviratne, Philip J. Ward, Andy Pitman, Amir AghaKouchak, David N. Bresch, Michael Leonard, Thomas Wahl, and Xuebin Zhang. Future climate risk from compound events. *Nature Climate Change*, 8(6):469–477, June 2018. ISSN 1758-6798. doi: 10.1038/s41558-018-0156-3.

Jakob Zscheischler, Olivia Martius, Seth Westra, Emanuele Bevacqua, Colin Raymond, Radley M. Horton, Bart van den Hurk, Amir AghaKouchak, Aglaé Jézéquel, Miguel D. Mahecha, Douglas Maraun, Alexandre M. Ramos, Nina N. Ridder, Wim Thiery, and Edoardo Vignotto. A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1(7):333–347, July 2020. ISSN 2662-138X. doi: 10.1038/s43017-020-0060-z.

APPENDIX

A HWB SCOPE AND BENCHMARK CONSTRUCTION

This appendix provides the benchmark composition, implementation details, baseline adaptations, evaluation protocol, and supplementary results supporting the main paper. All reported results use the same sealed 141-instance inventory and the DCA evaluator.

A.1 BENCHMARK SCOPE

HWB contains 141 sealed instances spanning seven single-hazard domains and four multi-hazard interaction classes. The final inventory contains 90 single-hazard and 51 multi-hazard instances. At the initial state, 45 instances admit one registered route and 96 expose multiple candidate routes.

The construction infrastructure begins from a global atlas of 3,062 registered geospatial scenarios, of which 542 were materialized for benchmark generation. The atlas is a construction resource: headline results are computed exclusively on the sealed 141-instance inventory. The selected instances preserve the diversity of their scientific sources while binding every task to an explicit output contract, candidate route set, capability state, and task-native evaluator.

The eleven source families provide observed events, forcing variables, operational products, and reference outputs for benchmark construction. They support wildfire, flood, landslide, earthquake, tropical-cyclone, drought, heatwave, and multi-hazard tasks across geographically diverse settings.

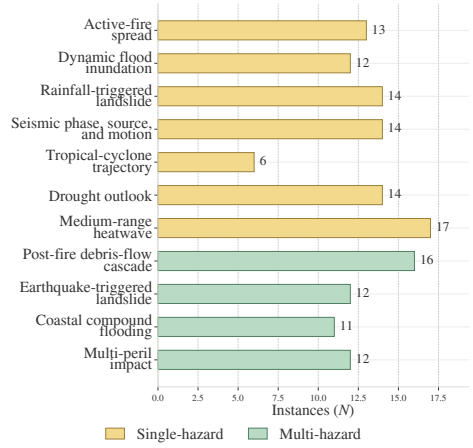


Figure A.1: HWB inventory. Instance counts across the 7 single- and 4 multi-hazard tracks.

A.2 HAZARD AND ROUTE STRATA

The hazard-composition and route-multiplicity strata are independent. Figure A.2 reports their joint distribution, allowing performance to be examined both by scientific composition and by the amount of route choice presented to the agent. Multiple-route cases occur in both single- and multi-hazard settings.

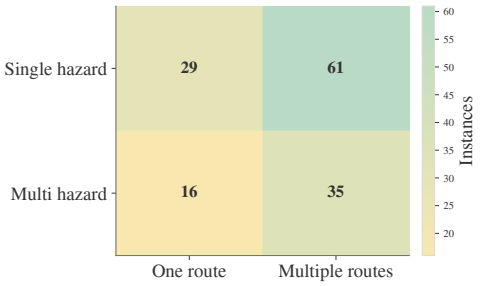


Figure A.2: Joint benchmark strata. Instance counts by hazard and route composition.

Each instance specifies a scientific target, initial state, information cutoff, candidate route set, capability state, output contract, and task-native evaluator. The 96 multiple-route instances expose more than one registered candidate at the initial state. The evaluator checks the committed route against the benchmark-side criterion bound to the corresponding scenario.

A.3 CONSTRUCTION AND SEALING

Benchmark construction proceeds through four stages. First, scientific datasets and operational products are mapped to task definitions with explicit output variables, units, spatial and temporal support, and evaluation criteria. Second, candidate route families are instantiated from Route Cards and executable capabilities. Third, every candidate instance undergoes schema, provenance, temporal-cutoff, and replay checks. Finally, the selected inventory, route registry, capability versions, evaluator contracts, and reference artifacts are frozen together.

Every sealed instance has at least one construction-time workflow that can produce the required artifact in a clean environment. This workflow establishes solvability during construction; it is not shown to the evaluated agent and does not prescribe its execution trajectory.

Table A.1: **Scientific data sources used in HWB construction.** Each source supplies observed events, forcing variables, reference products, or evaluator targets; the corresponding publications and official records are cited in Section 4.

Hazard task	Scientific source and role
Wildfire spread	**WildfireSpreadTS.** Time-indexed fire observations for evaluating the spatial and temporal progression of active-fire spread.
Flood inundation	**FloodCastBench.** Hydrological forcing fields and reference flood extents for dynamic inundation analysis.
Rainfall-triggered landslide	**NASA LHASA v2.** Global landslide nowcasts and reference occurrences for rainfall-triggered susceptibility analysis.
Earthquake analysis	**SeisBench.** Waveforms and reference labels for phase picking, event association, and seismic-source estimation.
Tropical-cyclone tracking	**TCBench.** Cyclone trajectories and lead-time-specific references for evaluating forecast track error.
Drought outlook	**NOAA CPC drought products.** Standardized precipitation indicators and operational references for drought assessment.
Heatwave forecasting	**ExtremeWeatherBench.** High-impact weather events and reference targets for medium-range forecast verification.
Post-fire debris flow	**USGS Post-Wildfire Debris-Flow Hazard Assessment Collection.** Burn severity, rainfall, likelihood, and volume for cascade analysis.
Earthquake-triggered ground failure	**USGS Ground Failure products.** Landslide and liquefaction estimates for ground-failure and coupled-impact assessment.
Coastal compound flooding	**SFINCS coastal simulations.** Inundation targets under interacting oceanic, pluvial, and fluvial forcing conditions.
Multi-peril impact	**NOAA Storm Events Database and HWB damage cases.** Event-level records for property-damage assessment across interacting hazards.

Table B.1: Principal Route Card fields.

Field group	Contents
Task identity	Hazard domain, task specification, requested output, and route family.
Scientific conditions	Assumptions, required observations, validity ranges, calibration regime, scale, and uncertainty treatment.
Capability requirements	Required datasets, models, formulas, transformations, and compatibility or dependency constraints.
Output contract	Variable definition, units, format, spatial and temporal support, resolution, and artifact schema.
Evaluation	Task-native metric, tolerance, and acceptance threshold.
Provenance	Scientific source, capability version, supporting evidence span, and benchmark-construction record.

B Route Cards, Capabilities, and Agent Operations

B.1 Route Card Representation

A Route Card records the scientific and executable information required to determine whether a route is applicable and realizable. Its principal fields are summarized in Table B.1. Full HazardWeaver resolves the frozen HKC-derived conditions associated with these Route Cards at the initial state and revisits their status when permitted evidence changes. The HKC-off condition in RQ2 removes this evidence-linked applicability layer while retaining the remainder of the evaluation pipeline.

Table B.2: Grounded tools available to HWA. Operations are instantiated only when permitted by the task and remaining execution state.

Operation	Purpose and state effect
Inspect state	Reads evidence, artifacts, and capability status; no state change.
Inspect route	Reads route conditions, capabilities, and contract; no state change.
Execute capability	Runs a capability; records artifact and provenance, or the failure.
Request information	Obtains a permitted unresolved observation; adds it as evidence.
Repair workflow	Inserts a transformation or subpath; keeps the route-level target.
Change route	Commits to another route from the recomputed eligible set.
Submit artifact	Returns the terminal artifact and invokes the sealed evaluator.
Submit abstention	Abstains after failed recovery; invokes the abstention criterion.

B.2 Typed Capability Representation

HCG contains 82 registered capability specifications corresponding to 79 distinct executable capabilities. Capabilities include scientific datasets, empirical equations, numerical models, learned models, simulators, transformations, and coupling operators. Each input and output port records the scientific variable, unit, representation, spatial and temporal support, resolution, uncertainty metadata, and provenance requirements. Representation-preserving transfers may connect capabilities directly. Operations that change information, including interpolation, aggregation, thresholding, learned downscaling, and unit conversion, remain explicit capabilities. Their effects therefore remain visible in the executed workflow and its provenance record.

B.3 Permitted Agent Operations

Table B.2 lists the grounded operations used by HWA. Every operation either inspects the current state, executes a registered capability, or produces an auditable state transition.

C Evaluation and Reporting Protocol

C.1 Decision-Constrained Accuracy

DCA follows Eq. 8 in the main paper. For an artifact submission, a pass requires task-native artifact success together with benchmark-side route validity, provenance validity, typed and causal execution validity, and the applicable resource contract. An abstention is credited only under the registered recovery criterion in Eq. 7. Missing, incomplete, or unevaluated submissions remain in the denominator and count as failures.

For reporting, E_q denotes the artifact-validity factor in Eq. 8, while V_q aggregates the corresponding trajectory-validity factors. Table 3 reports their marginal pass rates; headline DCA requires them jointly with benchmark-side route validity. Consequently, E_q or V_q may exceed DCA when the remaining criterion is not satisfied.

C.2 Task-Native Evaluators

HWB preserves the scientific evaluator appropriate to each output rather than forcing heterogeneous artifacts into a common numerical error. Table C.1 summarizes the evaluator families. Exact thresholds, output fields, and units are stored with the sealed instance contracts.

C.3 Statistical Reporting

Instance-level proportions use Wilson 95% confidence intervals. Macro DCA is the unweighted mean of the 11 track-level DCA rates; its 95% interval is obtained by resampling the 11 tracks with replacement for 5,000 bootstrap replicates using seed 42. Comparisons between HazardWeaver and individual baselines use an exact paired McNemar test over the same 141 sealed instances.

Table C.1: Task-native evaluator families. DCA applies the artifact criteria below jointly with route and execution validity.

Hazard track	Scientific outputs and evaluation
Active-fire spread	Fire probability, occupancy, or perimeter, evaluated using average precision, intersection over union, and registered field checks.
Dynamic flood inundation	Inundation depth and extent, evaluated using depth RMSE over the scenario-specific spatial support.
Rainfall-triggered landslide	Landslide occurrence probability, evaluated using area under the precision–recall curve.
Seismic phase, source, and motion	Phase picks, event associations, source parameters, and ground motion, evaluated using pick F1, association score, and coverage.
Tropical-cyclone trajectory	Forecast storm trajectories, evaluated using lead-time-specific track error in kilometers.
Drought outlook	Drought categories or probabilities, evaluated using the registered operational outlook-skill criterion.
Medium-range heatwave	Heatwave probabilities or intensities, evaluated against the maximum accepted forecast error.
Post-fire debris-flow cascade	Debris-flow likelihood, volume, or affected area, evaluated using output-specific volume error, Brier score, precision–recall.
Earthquake-triggered landslide	Triggered-landslide or service-impact outputs, evaluated through scientific replay under the registered coupling model.
Coastal compound flooding	Compound-flood depth and extent, evaluated using depth RMSE under the combined forcing conditions.
Multi-peril impact	Multi-peril loss or substrate coverage, evaluated using registered log-scale error and coverage criteria.

D Implementation Details

D.1 Language-Model Backbones

Experiments use seed 11, temperature 0.2, the frozen HKC registry, the matched execution protocol, and DCA rescoring. All compared methods operate on the same sealed inventory, candidate route catalog, scientific execution interface, evaluator contracts, and per-instance failure policy. A run that does not return a valid scored submission counts as a failure.

Table D.1: Policy backbones. Model identifiers used with Full HazardWeaver; all non-policy components remain fixed.

Backbone	Model identifier
OLMo-3.1-32B	allenai/Olmo-3.1-32B-Instruct
Mixtral-8x22B	mistralai/Mixtral-8x22B-Instruct-v0.1
Gemma-4-31B-it	google/gemma-4-31B-it
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct
DeepSeek-V4.1-Flash	deepseek-v4.1-flash

D.2 Execution Records

Each run stores the selected route, route commitments, capability invocations, generated artifacts, provenance edges, state transitions, and terminal submission. Numerical, spatial, and probabilistic outputs are generated by registered scientific capabilities; the policy model controls route selection and execution flow.

E Baseline Adaptations

The source systems target different execution environments and task types. We therefore adapt their published decision principles to the common HWB route-selection interface. Every method receives the same task description, candidate Route Cards, registered capability identifiers, public output

Table E.1: Baseline adaptations used in RQ1. "Rounds" refers to route-selection or reflection rounds before the terminal commitment.

Method	Adapted principle and HWB realization
AIDE	Iterative candidate refinement. Refines a proposed scientific route for up to three rounds using the task and Route Cards, then commits directly.
DS-Agent	Case-based retrieval. Retrieves the three most similar training-side route cases and uses them to rank the registered candidates.
R&D-Agent	Hypothesis-driven ordering. Forms a task hypothesis, orders the candidate scientific routes against it, and commits to the highest-ranked route.
MLE-STAR	Metric-directed program search. Infers the required scientific metric and output program before ranking candidate routes against those requirements.
AutoML-Agent	Pipeline and schema selection. Infers the task input–output schema and ranks routes by their ability to realize a compatible scientific pipeline.
DisasterBench-ToT	Deliberate search over disaster tools. Treats registered capabilities as typed disaster tools and performs bounded tree search over candidate route choices.
ReAct	Interleaved reasoning and action. Alternates route inspection and comparison for up to three turns before issuing a final commitment.
Reflexion	Verbal reflection from execution outcomes. Revises its route choice for up to three rounds using a qualitative failure class from the preceding execution.
Self-Consistency	Sampling and majority agreement. Produces five independent route rankings and commits to the candidate receiving the largest number of first-place votes.
Plan-and-Solve	Explicit planning before execution. Produces an ordered scientific plan and executes its first committed route under the shared interface.

requirements, and scientific execution interface. After route selection, the chosen route is executed and evaluated through the same sealed replay and DCA stack used for HazardWeaver.

The adaptations preserve the characteristic control strategy of each method while avoiding method-specific scientific tools, private reference artifacts, or evaluator feedback during route selection. Table E.1 summarizes the resulting implementations.

These adaptations separate route-selection strategy from scientific execution. Their outputs remain fully comparable because the selected capabilities are evaluated on the same instances with the same artifact and route contracts.

F Supplementary Experimental Results

F.1 Exact RQ2 Ablation Results

Table F.1 gives the counts underlying Figure 3. The 96 instances are the multiple-route stratum of the sealed 141-instance benchmark. Removing either evidence-linked applicability conditions or typed capability composition reduces DCA for every backbone.

Table F.1: Scientific-constraint ablations. DCA on the 96 route-choice instances; entries give count and percentage.

Backbone	HKC off	HCG untyped	Full
OLMo	38 (39.6)	39 (40.6)	**49 (51.0)**
Mixtral	34 (35.4)	32 (33.3)	**65 (67.7)**
Gemma	50 (52.1)	55 (57.3)	**76 (79.2)**
DeepSeek	68 (70.8)	71 (74.0)	**86 (89.6)**
Llama	44 (45.8)	46 (47.9)	**90 (93.8)**

The route-update ablation in Table 2 uses 40 applicable instances drawn from the sealed pool. Across five backbones, updated eligibility passes 83 of 200 evaluated backbone–cell combinations, compared with 32 of 200 under fixed initial eligibility. This pooled value is descriptive; the per-backbone comparisons remain the primary result.

F.2 Paired Significance Tests

The test is paired at the instance level: for each baseline, HWA-only counts cases solved exclusively by HazardWeaver, whereas BL-only counts cases solved exclusively by that baseline. This construction controls for variation in task difficulty because both systems are evaluated on exactly the same 141 instances. The imbalance is substantial for every comparison. Even against Plan-and-Solve, the strongest baseline by aggregate DCA, HazardWeaver uniquely solves 75 instances while losing only seven in the opposite direction. The pattern becomes still more pronounced for the shared-tool controllers; for example, HazardWeaver records 84 unique successes against ReAct with no ReAct-only success. All exact McNemar tests remain significant by a wide margin.

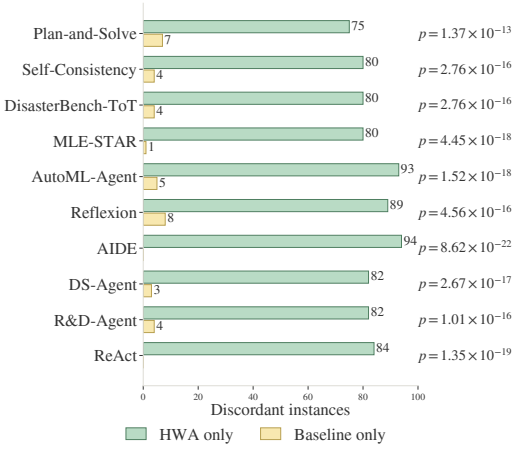


Figure F.1: Paired significance tests. Discordant DCA counts between HazardWeaver and each baseline; annotations give McNemar p-values.

These paired results complement the aggregate comparison in Table 1. The advantage is not produced by averaging over a small number of unusually favorable tracks or by one isolated family of tasks. Instead, the discordant outcomes consistently favor HazardWeaver across the sealed inventory and across baseline designs based on planning, reflection, retrieval, iterative refinement, and sampled agreement. This provides instance-level evidence that explicit scientific route eligibility improves the reliability of the resulting decisions beyond the particular control strategy used for comparison.

F.3 Joint Hazard and Route Strata

Figure F.2 separates hazard composition from route multiplicity instead of treating either property as a proxy for the other. HazardWeaver performs strongly in both route-rich strata, solving 58 of 61 single-hazard and 32 of 35 multi-hazard instances with multiple candidate routes. Its advantage is therefore preserved when route choice is combined with a composed scientific analysis. Plan-and-Solve is competitive on the smaller multi-hazard, one-route subset, where it solves 13 of 16 instances compared with 12 for HazardWeaver. This local result also shows that the overall comparison is not uniformly favorable by construction. The largest separation appears once the task requires choosing among several routes. Plan-and-Solve falls from 13 of 16 on multi-hazard, one-route tasks to eight of 35 on the corresponding multiple-route stratum, whereas HazardWeaver solves 32 of those 35 instances. AIDE shows a similar difference between the route regimes. Together with the controlled ablations in RQ2, this stratification supports that HazardWeaver's principal advantage lies in organizing scientific route choice while retaining strong execution across both isolated and composed hazard analyses.

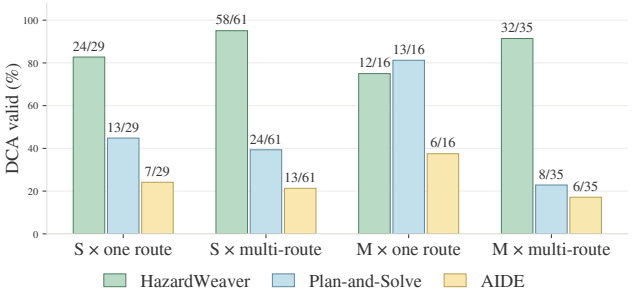


Figure F.2: Joint-stratum performance. DCA across hazard composition and route multiplicity; annotations report successful instances over stratum size.

F.4 Track-Level Cross-Backbone Results

Performance is distributed across all 11 tracks. The tropical-cyclone trajectory track is solved by every backbone, while the dynamic flood inundation track remains a shared bottleneck. The multi-hazard tracks produce greater cross-backbone separation, consistent with the broader profiles shown in Figure 4c.

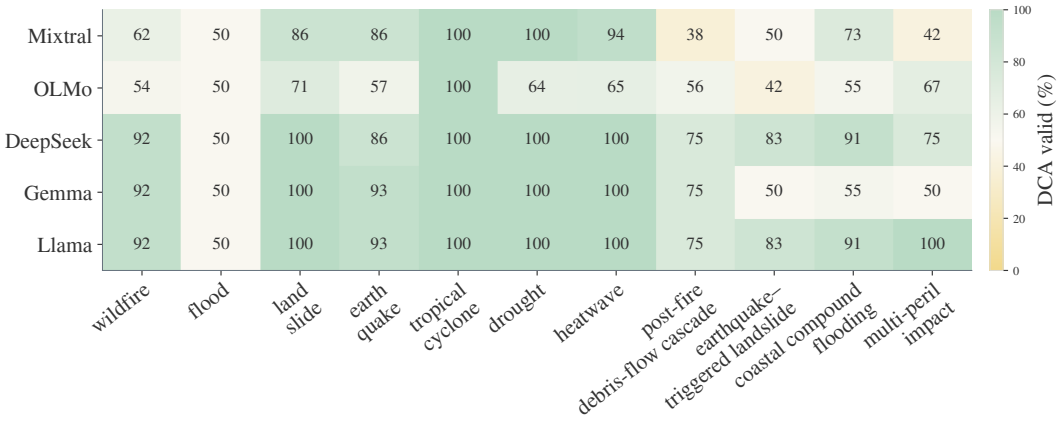


Figure F.3: Track-level DCA for Full HazardWeaver. Each entry reports track-level DCA (%).

G Benchmark Sealing and Quality Control

G.1 Evaluator Validation

Before sealing, the evaluator was exercised on 160 controlled fixtures spanning eight decision classes: valid alternative execution, scientifically inapplicable route, capability-incompatible route, justified abstention, unjustified abstention, post-cutoff information use, prediction without capability provenance, and coupling-insensitive multi-hazard output. All 160 fixtures produced their expected accept or reject decision.

G.2 Information Separation

Table G.1: **Information separation.** Availability for execution and scoring.

Information	Agent	Eval.
Task and initial state	✓	✓
Candidate routes	✓	✓
Capability interface	✓	✓
Public output contract	✓	✓
Reference artifact		✓
Sealed route criterion		✓
Construction workflow		✓

Reference artifacts, evaluator thresholds beyond the public task specification, construction-time workflows, and benchmark-side route judgments are held outside the solver context. Agents receive the task, initial state, candidate route descriptions, permitted capabilities, and public output requirements. Evaluation is performed only after terminal submission. All benchmark-side route judgments are frozen before any evaluated policy run; evaluation never derives or revises its route criterion from the tested system's output.

G.3 Route-Registry Review

We conducted a stratified human review of 86 route–condition pairs drawn from 44 benchmark cells across all 11 tracks. All 27 sampled witness routes were judged scientifically applicable. Candidates rejected by the sealed contract were subsequently adjudicated against the requested output, scientific semantics, and capability requirements; this review found no route that was both contract-valid and omitted from the registry. We use this review as quality control for the sealing procedure rather than as a claim that the catalog exhausts every route that future scientific work may establish.

H Representative Execution Traces

H.1 Route Revision After an Execution-State Change

In a representative seismic instance, HWA initially enumerates the available seismic routes, commits to the capability satisfying the phase-picking contract, and begins execution. An intermediate state change invalidates the current realization. HWA updates the capability state, reinstantiates a compatible realization, recommits under the revised state, and completes the required artifact. The final route and phase-picking output both satisfy the sealed evaluator.

Table I.1: Reproducibility artifacts accompanying HWB.

Artifact	Contents
Sealed inventory	The 141 instance identifiers, track assignments, task specifications, initial states, route strata, and evaluator bindings.
Route Cards	Scientific conditions, required observations, capability signatures, output contracts, uncertainty fields, and evidence provenance.
Capability registry	Executable entry points, versioned dependencies, typed input and output ports, transformations, and validation metadata.
Evaluator contracts	Task-native metrics, tolerances, route criteria, cutoff rules, provenance requirements, and abstention conditions.
Run configurations	Model identifiers, stochastic settings, execution protocol, and method-specific selector configuration.
Per-instance outputs	Selected routes, state transitions, capability invocations, generated artifacts, provenance records, and terminal evaluation reports.
Aggregation scripts	DCA, marginal validity, macro DCA, Wilson intervals, bootstrap intervals, and paired significance calculations.
Figure and table inputs	Machine-readable values used to generate every experimental figure and table in the paper.

This trace illustrates the distinction between retaining a scientific route family and repairing its executable realization: the scientific target remains appropriate, while the capability path must be reconstructed from the updated state.

H.2 Failure Under an Incomplete Cascade

A representative post-fire debris-flow cascade failure occurs when the policy selects a debris-flow volume route for a task requiring the complete burn–rainfall–debris-flow cascade. The selected capability produces a scientifically meaningful intermediate quantity but does not satisfy the terminal route and output contract. The evaluator therefore rejects the submission. This case shows that successful execution of an individual scientific model does not by itself establish a decision-valid multi-stage analysis.

I Reproducibility and Release Contents

The release is organized around the sealed inventory so that every reported aggregate can be reconstructed from per-instance records. Table I.1 summarizes the principal artifacts.

A complete reproduction first materializes the frozen task and capability versions, executes the selected system configuration on all 141 instances, applies the sealed evaluator to each terminal record, and then aggregates the resulting per-instance decisions. The released table inputs provide an additional direct check that the manuscript values match the sealed evaluation outputs.

J Limitations

HazardWeaver operates within a frozen scientific capability catalog and evidence index. HKC conditions reflect the literature corpus and Route Cards available at benchmark construction; new evidence, revised standards, or updated models may require recompilation and review. HWB evaluates artifacts and committed routes through frozen task-specific contracts, so reported validity denotes reproducible agreement with these documented criteria. Headline scoring covers routes registered when the benchmark is sealed; newly established routes require catalog extension and resealing. The 141 instances span eleven hazard tracks and diverse geographic settings without exhausting every event, region, sensor configuration, or operational constraint. Operational forecasting and emergency decision-making therefore require updated evidence, local calibration, and domain review.