

Article

# Separating Runtime Acceptance from Offline Contract Alignment in LLM-Orchestrated Flood GIS Workflows: A Dual-Case Evaluation

Hongyun Zhang <sup>1,\*</sup> , Jin Liu <sup>2</sup> , Fang Wang <sup>1</sup> , Yahong Zhao <sup>1</sup> and Ye Xuan <sup>3</sup>

<sup>1</sup> School of Disaster Prevention and Mitigation Engineering, University of Emergency Management, Langfang 065201, China; wangfang@ncist.edu.cn (F.W.); neu\_zyh@163.com (Y.Z.)

<sup>2</sup> State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; jliu@sgg.whu.edu.cn

<sup>3</sup> School of Geomatics, Liaoning Technical University, Fuxin 123000, China; 47231056@stu.lntu.edu.cn

\* Correspondence: zhanghongyun@whu.edu.cn; Tel.: +86-159-0423-4253

## Abstract

Flood assessments require reproducible links between spatial products and the measurements ultimately reported to analysts, yet a large language model (LLM) agent may complete a geographic information system (GIS) workflow without generating the required product or traceable claim. We present GeoFloodAgent, an LLM-orchestrated flood-GIS workflow that combines typed planning and deterministic GIS tools with a separate evaluation of runtime disposition and post-run alignment with task-specific product, lineage, and action contracts. We evaluated 80 Chinese-language tasks from the Zhengzhou (2021) and Forlì (2023) flood cases under seven configurations and three repetitions, yielding a registry-remediated composite of 1680 assignments. In the Full configuration, 212 assignments (88.3%) were accepted and contract-aligned, whereas 16 accepted assignments did not align with their registered contracts; all 16 concerned clarification or refusal actions. Three runs were rejected despite retained contract-aligned evidence. The cluster-weighted Full-Tool Agent differences were −1.15 percentage points for accepted contract mismatch and +1.07 percentage points for contract-aligned acceptance, with empirical intervals spanning benefit and harm. Under frozen Zhengzhou inputs, deterministic outputs were sensitive to the minimum SAR object-size setting: relative to 80 pixels, candidate flood extent changed by +105.6% at 40 pixels and −30.9% at 120 pixels, whereas a 20 m simplification tolerance reproducibly failed during geometric union. These configuration-specific results do not validate hydrological accuracy or identify universal optimal defaults. These findings show that completion rate alone is inadequate for evaluating LLM-orchestrated flood GIS: runtime disposition and evidence alignment should be reported jointly. The study evaluates a finite, study-authored benchmark and does not establish hydrological accuracy, operational decision validity, or open-world error detection.

**Keywords:** large language models; geographic information systems; flood



Academic Editors: Wolfgang Kainz and Godwin Yeboah

Received: 12 August 2026

Revised: 21 September 2026

Accepted: 25 September 2026

Published: 1 October 2026

**Copyright:** © 2026 by the authors.

Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Flood assessment commonly combines satellite observations, rainfall and terrain information, infrastructure inventories, and GIS overlay analysis to produce inundation, exposure, and risk products for analyst review. These workflows remain dependent on explicit choices about input data, coordinate reference systems, spatial processing, and the

evidential basis of reported measurements. Satellite Earth observation has expanded flood mapping and monitoring capabilities [1–3], while GIS-based vulnerability and susceptibility studies show the continuing importance of spatial context in flood-risk analysis [4,5]. Artificial intelligence can assist with integrating heterogeneous information and organizing analytical procedures, but it does not remove the need to verify the geographic products on which flood-related claims are based.

Recent LLM-enabled GIS studies show how natural-language requests can be translated into tool calls, code, structured spatial workflows, and GIS outputs [6–8]. Related work also applies large vision-language models to post-disaster damage assessment and reporting [9]. These advances are valuable, but a completed trajectory or a plausible report does not by itself establish that a required flood layer exists, that a geometry is structurally usable, or that a reported area, length, count, or action can be traced to an admissible product. Conversely, a runtime check may reject a trajectory even when the archived products needed to satisfy the task are already present. A single completion metric cannot distinguish these outcomes.

This study therefore evaluates GeoFloodAgent as an intermediate assurance layer for LLM-orchestrated flood-GIS workflows. A hosted LLM produces a typed plan, registered deterministic GIS tools create spatial products and measurements, and a metric-level evidence ledger records their declared lineage. A runtime validator checks computable product, geometry, archive-profile, and lineage requirements, while a separate post-run evaluator compares serialized observations with frozen product and action contracts. Expected values and numerical tolerances are withheld from planner prompts and runtime-validator interfaces.

The framework is evaluated with 80 template-structured Chinese-language tasks derived from the Zhengzhou and Forlì flood cases. Zhengzhou represents an end-to-end SAR-to-report setting, whereas Forlì begins with a prepared Copernicus Emergency Management Service (CEMS) observed extent and tests downstream use of an existing operational product. The study reports the four mutually exclusive runtime-disposition  $\times$  contract-alignment outcomes and compares the Full configuration with a same-endpoint Tool Agent as an exploratory system-level contrast.

The central contribution is a task-specific evaluation design that separates runtime acceptance from post-run agreement with registered flood products, metric lineage, and boundary actions. Rather than claiming a more accurate flood map or an autonomous emergency-decision system, the study tests whether an LLM-orchestrated GIS workflow leaves auditable evidence for the products and claims it is asked to produce.

The revised paper makes two focused contributions: it operationalizes a product- and evidence-conformance layer for LLM-orchestrated flood-GIS tasks, and it evaluates that layer across two cases with explicitly different data roles. Detailed implementation choices, ancillary diagnostics, and limitations are reported in the Methods and Discussion rather than repeated as contributions.

## 2. Related Work

### 2.1. LLM-Enabled GIS Workflow Generation

Recent work has moved from free-form code generation toward more structured GIS workflows. ChatGeoAI maps recognized entities and ontology terms to PyQGIS code [6], while a later framework builds an intermediate directed acyclic graph and revises GeoPandas and Shapely operations from execution feedback [8]. A complementary requirement is provenance representation: the W3C PROV Ontology (PROV-O) provides a vocabulary for describing entities, activities, use, and derivation, thereby supporting explicit links among source data, processing steps, and derived GIS products [10]. MeteoChat decouples an-

alytical logic from environmental data sources for reusable reporting [11]. Despite these architectural and provenance-oriented advances, the reported assessments focus mainly on executable code, completed workflows, or generated reports; they do not necessarily test whether a task-required product exists or whether a reported metric can be traced to an admissible artifact.

Tool-augmented agents further constrain how models interact with computational environments. ReAct interleaves reasoning with actions [12], Toolformer learns when to invoke external tools [13], and the Autonomous GIS vision extends these ideas to spatial analysis [14]. Within GIS, ShapefileGPT separates planning from execution over a documented vector-processing library [7], whereas GeoAgentBench evaluates agents in a sandbox of 117 atomic tools and includes multimodal spatial outputs [15]. Related systems address geospatial data acquisition [16], GeoSPARQL generation [17], and alignment between authoritative taxonomies and collaborative folksonomies [18]. These advances make tool choice and execution more observable, but they do not explicitly distinguish runtime acceptance from post-execution agreement with task-specific product and claim requirements.

## 2.2. LLMs in Flood and Disaster Workflows

Conventional flood-GIS research integrates remotely sensed observations, precipitation and terrain variables, exposure overlays, spatial indicators, and data-driven models to characterize flood hazards and their consequences [1–5]. LLM orchestration is complementary to these approaches: it may coordinate established analysis steps, but it does not by itself validate the hydrological interpretation or decision relevance of their outputs.

Flood and disaster workflows extend the same problem over longer trajectories, where early mistakes can alter downstream products. A knowledge-constrained flood system combines GIS and flood knowledge graphs, retrieval, ArcPy execution, and textual descriptions of spatial outputs [19]. DisasTeller coordinates specialized agents across damage assessment, alerts, resource allocation, and recovery reporting, and shows how early errors can propagate to downstream products [9]. DORA extends evaluation to 515 tasks from 45 events and 108 geospatial tools, identifying persistent bottlenecks in tool selection and argument grounding as trajectories lengthen [20]. These studies demonstrate the importance of domain grounding and broad tool coverage, but do not examine whether an accepted flood-GIS run leaves behind the products and traceable claims specified by the task.

## 2.3. Spatial Product Quality, Provenance, and Contracts

The evaluation combines ideas from Design by Contract, geographic data-quality standards, and provenance modeling. Design by Contract expresses obligations as preconditions, postconditions, and invariants [21]. ISO 19157-1 treats geographic data quality as conformity to a product specification, while the Open Geospatial Consortium (OGC) Simple Feature Access standard specifies geometry semantics [22,23]. The World Wide Web Consortium (W3C) PROV Ontology (PROV-O) represents entities, activities, use, and derivation [10], and subsequent workflow research applies related models to reusable projects, OGC Web Processing Service workflows, and semantic service composition [24–26]. The SpatioTemporal Asset Catalog (STAC) provides a complementary model for discovering and linking spatiotemporal assets [27].

These standards do not themselves define an acceptance rule for a particular flood task. GeoFloodAgent therefore uses a narrower runtime ledger that binds each reported metric to the tool version, parameters, source layers, CRS metadata, and product generated in the same run. A registered contract can then ask whether a required product exists

and whether a claim is supported by an admissible source. The ledger operates alongside PROV-O, STAC, and OGC standards rather than serving as an alternative to them.

Against this background, recent LLM applications in geospatial reporting, geospatial question answering, semantic alignment, and location extraction [11,17,18,28,29] are generally assessed through the correctness, usefulness, or completion of the generated answer or task output. GeoFloodAgent instead evaluates runtime disposition jointly with frozen product, lineage, and action contracts across two flood cases. Table 1 summarizes this positioning.

**Table 1.** Positioning of GeoFloodAgent relative to representative studies of large language models (LLMs) for geographic information systems (GIS).

| Research Line                               | Primary Unit                                                                 | Typical Strength                                                            | Distinction Emphasized in the Present Study                                          |
|---------------------------------------------|------------------------------------------------------------------------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Natural language to GIS code/workflow [6,8] | code or workflow completion                                                  | broad user access and executable analysis                                   | offline checks of required layers, geometry, actions, and claim lineage              |
| Specialized GIS agents [7,16]               | function call or acquired data                                               | constrained tool use and modularity                                         | whether a reported flood metric still points to the spatial product that produced it |
| GeoAgentBench [15]                          | executed trajectory and spatial output                                       | dynamic GIS-agent sandbox                                                   | broader GIS benchmark; no numerical comparison with the present study.               |
| DORA [20]                                   | heterogeneous disaster trajectory                                            | multi-event disaster-agent evaluation                                       | broader event scope; the present study adds post-run contract alignment.             |
| Provenance standards [10,27]                | interoperable entities, activities and assets                                | general lineage and discovery models                                        | product-level checks and bounded repair at the LLM–GIS boundary for flood workflows  |
| GeoFloodAgent                               | declared flood spatial products, measurements and frozen output observations | joint evaluation of spatial products, metric lineage, and runtime decisions | runtime disposition separated from product-, claim-, and action-contract alignment   |

### 3. Materials and Methods

#### 3.1. Study Objective and Evaluation Boundary

We evaluated whether LLM-orchestrated flood-GIS workflows returned the products and traceable metric evidence specified for each task. The benchmark comprised 70 artifact-bearing contracts and 10 action contracts. Each artifact-bearing contract specified required products, one registered claim, a unit where applicable, and admissible evidence; each action contract required either clarification or refusal. Runtime disposition was evaluated separately from offline contract alignment, defined as agreement between the captured output and its frozen contract. A runtime-accepted but non-aligned output was classified as an accepted contract mismatch.

GeoFloodAgent separates five responsibilities. A hosted LLM proposes a structured tool graph; registered GIS tools perform numerical and spatial operations; an evidence ledger records metric-level lineage; a runtime validator checks computable product, geometry, and lineage properties; and a separate archive-evaluation script scores serialized observations after loading the expected values. Figure 1 shows these responsibilities and locates the implemented checks within a broader GIS assurance ladder.

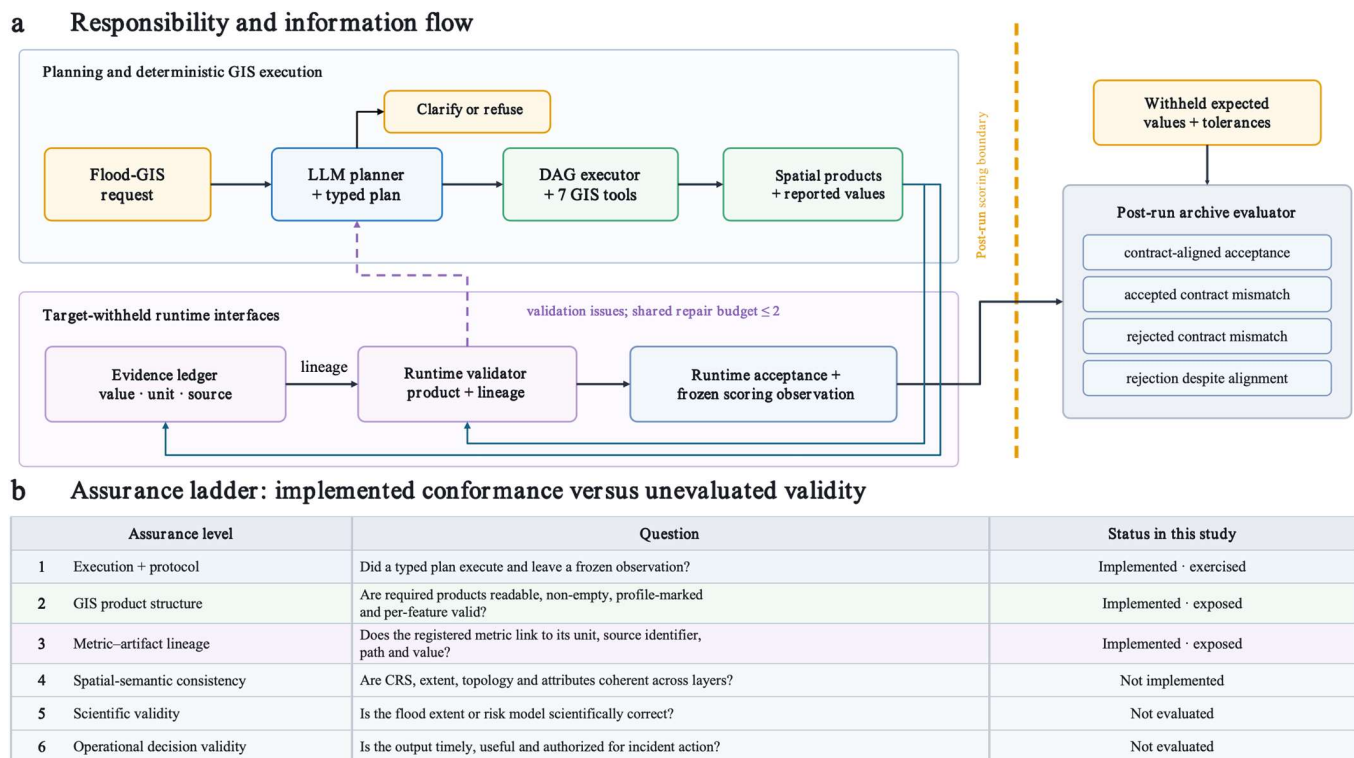


**Figure 1.** GeoFloodAgent architecture and assurance boundary. (a) A large language model (LLM) planner produces a typed directed acyclic graph (DAG); deterministic geographic information system (GIS) tools create products and measurements; and an evidence ledger links metrics to artifacts. Runtime validation precedes separate post-run archive evaluation, and the dashed arrow denotes bounded repair. (b) Implemented checks cover execution state, selected product structure, and registered metric lineage; cross-layer semantics, scientific validity, and operational validity remain outside scope. The legacy top-level crs field is an archive-profile marker, not a coordinate reference system (CRS) correctness test.

The contracts operationalize product structure, geometry validity, archive-profile metadata, units, and metric-artifact lineage. Because the benchmark assets were developed within this study and derive from two events, inference is confined to the study-authored specification and the finite benchmark described in Sections 3.4 and 3.9.

### 3.2. GeoFloodAgent Architecture

Figure 1a shows the runtime path from natural-language request to spatial products and the post-run archive-evaluation path; Figure 1b locates the implemented checks within a broader assurance ladder.

#### 3.2.1. Structured Planning and the Execution Boundary

For each task, the planner receives the Chinese prompt together with task and case identifiers, event metadata, available layers, the seven-tool catalog, tool dependencies, declared artifacts, and the two boundary actions. Field-level schemas for the plan, tool inputs, and boundary-action arguments are supplied in every condition except Schema Metadata Withheld. The instruction requests the smallest sufficient directed acyclic graph (DAG) using only registered tools, layers, measurements, and case data.

Each model response was required to contain a single JavaScript Object Notation (JSON) plan, which was parsed into a typed plan representation containing the task identifier, case identifier, and an ordered sequence of steps. Before execution, the system verified that every selected tool was registered, each argument satisfied its declared input schema,

the task and case identifiers were unchanged, and the step dependencies formed a directed acyclic graph with all mandatory upstream operations present.

The system also supported two boundary actions for requests that could not be safely converted into an executable GIS workflow. A clarification action was required when the case context or a necessary parameter was missing or conflicting, whereas a refusal action was required for unsupported predictions, unjustified certainty claims, or autonomous high-stakes decisions. A boundary action was terminal: when selected, it was the only step in the plan, and no GIS tool was executed.

To make the natural-language interface inspectable, Appendix A reproduces two representative Chinese-language prompts: one artifact-bearing task and one capability-boundary task. For each example, the appendix provides an English gloss, the declared case and layer context, and the required structured response fields. The complete frozen task registry, bilingual prompt fields, contract definitions, plan schemas, execution manifests, derived evaluation records, and analysis scripts are openly available in the versioned repository identified in the Data Availability Statement. These examples document the tested interface; they are not presented as representative records of real incident-command requests.

### 3.2.2. Deterministic GIS Tools

All workflow-generated spatial and numerical products were created by versioned Python tools (Python 3.12.10 for the original campaign and Python 3.12.4 for the registry-remediation campaign) rather than model-generated code. Prepared case inputs were registered separately and were not attributed to GeoFloodAgent. Each tool has a typed input/output model, declared dependencies, required layer identifiers, and an artifact contract. The core environment used GeoPandas 1.0.1, Shapely 2.0.6, Rasterio 1.4.3, and xarray 2024.11.0 [30–33].

Unless otherwise noted, the deterministic workflow used an 80-pixel minimum SAR candidate-object size, a 200 m facility-proximity buffer, a 10 m vector-simplification tolerance, and a 1000 m risk-grid cell. The first three settings were selected before benchmark construction as operational processing defaults rather than calibrated hydrological thresholds. At the 10 m SAR grid resolution, 80 pixels correspond to 0.008 km<sup>2</sup> and remove isolated candidate components before vectorization. Connected-object filtering is commonly used to suppress isolated SAR detections [34,35]. The 200 m buffer is used only to screen facilities in proximity to a candidate flood polygon, consistent with transparent exposure-screening practice [36]; it is not interpreted as a physical impact zone or evacuation distance. The 10 m tolerance is a cartographic simplification setting [37], and its influence on downstream summary products is examined separately below. For Zhengzhou, Sentinel-1 C-band ground range detected (GRD) scenes were obtained from Google Earth Engine [38] (collection 'COPERNICUS/S1\_GRD') and median-composited within each acquisition window. Scenes were filtered to interferometric wide (IW) swath mode and 10 m resolution, and vertical-transmit/vertical-receive (VV) and vertical-transmit/horizontal-receive (VH) polarizations were selected; orbit direction and relative orbit were not filtered.

The extractor calculated post-minus-pre differences. Candidate pixels had to meet four conditions: all four backscatter inputs were finite, post-event VH was below an Otsu threshold [39], the VH difference was below its empirical 20th percentile, and the European Commission Joint Research Centre (JRC) Global Surface Water occurrence layer, thresholded at 80%, did not mark the pixel [40]. The Otsu threshold defaulted to −20 dB when fewer than 100 finite values were available or estimation failed; the difference threshold defaulted to −2 dB only when no valid pixel was available. Objects smaller than 80 pixels were removed before vectorization. VV differences were retained as diagnostic artifacts

but did not enter the classification predicate. We interpret the resulting layer as a reproducible change-detection candidate rather than a calibrated inundation map.

Within each case, flooded area, affected-road length, and critical-facility count were divided by their respective maxima across the retained risk-grid cells. If a component maximum was non-positive, that component contributed zero to every cell. The normalized components contributed up to 50, 30, and 20 points, respectively, and the summed score was clipped to [0, 100]. Scores were classified as low [0, 25], medium (25, 45], and high (45, 100].

Area and length metrics were computed after reprojection to a locally appropriate projected CRS. The local Universal Transverse Mercator (UTM) CRS was estimated from the spatial extent of each analysis area using GeoPandas; this procedure selected EPSG:32649 for Zhengzhou and EPSG:32633 for Forlì. EPSG:6933 was used as a predefined fallback for wider extents. Areas and lengths were calculated in projected meters and converted to square kilometers or kilometers before serialization and ledger registration.

A one-factor-at-a-time deterministic GIS sensitivity analysis was conducted using frozen case inputs. In Zhengzhou, the minimum SAR candidate-object size was varied over 40/80/120 pixels, the facility-proximity buffer over 100/200/300 m, and the vector-simplification tolerance over 0/10/20 m. In Forlì, which begins from a prepared EMSR664 observed extent, the minimum SAR object-size setting was not applicable; the buffer and simplification settings were varied over the same ranges. All non-varied inputs and processing settings were held fixed. We recorded candidate flood extent, polygon count, flood-intersecting road length, facilities within the buffer, and available risk-grid summaries relative to the case-specific baseline. No LLM call, benchmark assignment, or primary runtime/contract-alignment endpoint was recomputed. The analysis characterizes deterministic processing sensitivity, not hydrological accuracy.

### 3.2.3. Metric-Level Evidence Ledger

After each successful tool run, the evidence component flattens scalar fields from the canonical summary artifact into metric-level ledger records. Each row contains the task and plan step, tool identifier and version, metric path, value, unit when declared, CRS, source artifact identifier and path, required source layers, and executed parameters. Because the records derive directly from the execution trace and spatial products, the LLM does not author them. The ledger provides study-specific product-level traceability; it is separate from formal implementations of W3C PROV, OGC processing standards, or STAC.

### 3.2.4. Runtime Geospatial and Lineage Contracts

For validator-enabled conditions, the runtime receives required artifact identifiers, kinds and required/optional status. For claims tied to artifacts, it also receives a target-withheld evidence requirement containing only claim identifier, metric path, unit when applicable, and permitted evidence-artifact identifiers. Expected values, numeric tolerances and gold tool trajectories are omitted from planner prompts and the runtime-validator interface.

The runtime validator assesses task-specific product or action requirements, while offline alignment is evaluated separately. ISO 19157-1:2023 provides the broader product-specification view of geographic-data quality, and OGC Simple Feature Access provides the geometry concepts used here [22,23].

Vector validation follows a project-specific legacy archive profile used by downstream GIS scripts. The profile requires a literal top-level crs member, a non-empty features array, and per-feature planar simple-feature geometries accepted by Shapely. Exposure and risk tools serialize vectors in EPSG:4326, whereas SAR-derived polygons retain the reference-

raster CRS; metric operations use a local projected CRS. This is an interoperability requirement rather than a compliance test for Request for Comments (RFC) 7946. RFC 7946 fixes GeoJSON coordinates to World Geodetic System 1984 (WGS 84)/OGC CRS84 and assigns no normative CRS semantics to foreign members [41]. The OGC Features and Geometries JSON (JSON-FG) standard provides a standards-oriented option for archives that require explicit CRS semantics [42].

The validator also checks artifact availability and readability; GeoTIFF CRS, dimensions, and valid pixels; ledger representation of the required metric and, where applicable, its unit; use of a permitted evidence artifact; and agreement between ledger values and their resolved JSON or GeoJSON sources. Together, these checks govern flood, exposure, and risk products and the claims linked to them. Table 2 distinguishes runtime conformance, post-run archive assessment, and scientific questions outside both layers.

**Table 2.** Registered geographic information system (GIS) product and capability-boundary contract objects.

| Contract Object                             | GIS or Operational Meaning                                            | Runtime Check                                                                                           | Post-Run Archive Check                                                                                                | Outside the Evaluated Contract                                                                                                |
|---------------------------------------------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Required spatial product or report artifact | A declared flood, exposure, risk, context, or reporting product       | Identifier resolves; file is readable, non-empty, and structurally consistent with its kind             | The captured required artifact passes the kind-specific structural check; content and checksum are retained for audit | Hydrological correctness, model optimality, and completeness of unrestricted report prose                                     |
| GeoJSON geometry structure                  | A usable non-empty vector product                                     | Parseable root, non-empty features, and planar simple-feature validity                                  | Captured root type, non-empty feature count, profile-marker presence, and valid-geometry count pass structural checks | Null-feature completeness, cross-layer topology, attribute semantics, and hydrological plausibility                           |
| Project archive-profile marker              | A project-specific serialization convention                           | Legacy top-level crs member for vectors; readable CRS for GeoTIFF                                       | Presence of the registered vector-profile marker                                                                      | RFC 7946 compliance, wrong-but-present CRS labels, CRS correctness, axis order, coordinate range, and cross-layer consistency |
| Metric-artifact lineage                     | The product supporting an area, length, count, or risk value          | Metric path and unit resolve to a permitted source artifact; JSON/GeoJSON values agree with that source | One registered value extracted from a permitted captured product matches the target within tolerance                  | Scientific validity, causal interpretation, and decision utility                                                              |
| Capability-boundary action                  | Clarification or refusal for an underspecified or unsupported request | A valid boundary action is recorded as the sole plan step                                               | Observed action matches the registered clarification or refusal                                                       | Explanation quality and real-world operator utility                                                                           |

Abbreviations: CRS, coordinate reference system; JSON, JavaScript Object Notation; RFC, Request for Comments.

Runtime checks in Table 2 apply to validator-enabled paths; the same post-run archive evaluator was applied to every condition. Expected values and tolerances were omitted from planner prompts and runtime-validator interfaces, then loaded by the archive evaluator for primary scoring.

### 3.2.5. Bounded Repair

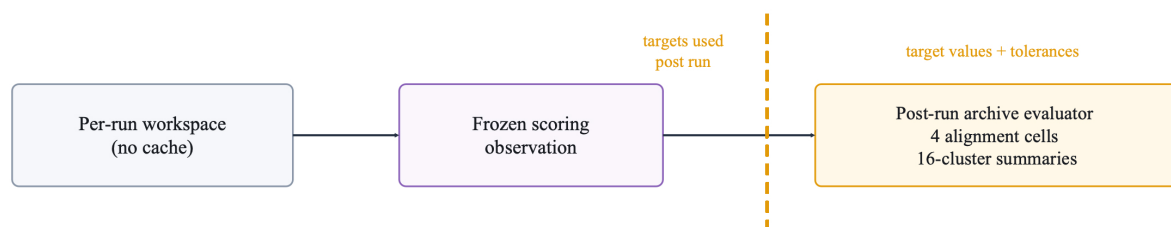
Schema and execution validation share a maximum budget of two repair attempts. Pre-execution errors and validator failures marked as repairable return issue codes and affected artifact identifiers, but never gold values, to the same planner, which must produce a corrected minimal plan. Persistent and non-repairable failures remain explicit.

Figure 2 summarizes the experimental conditions and registry-remediated composite described in Sections 3.5 and 3.6, together with the target-withholding scheme described in Section 3.7.

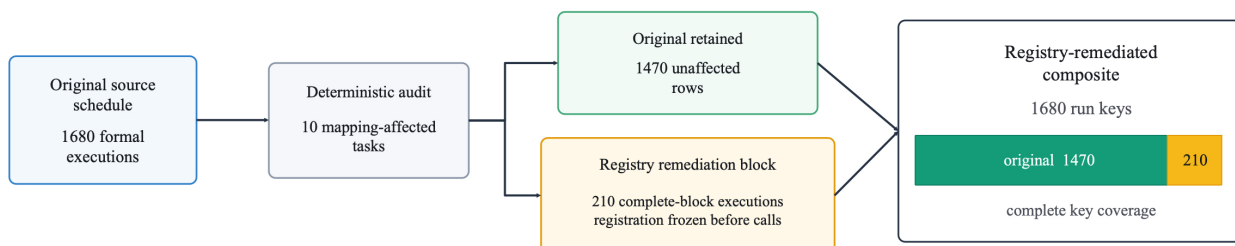
#### a Core design and post-run scoring

80 frozen tasks  $\times$  3 repetitions  $\times$  7 conditions = 1680 assigned runs

240 task-repetition blocks; condition order randomized within each source schedule



#### b Registry remediation and source provenance



Affected original rows (210): excluded from the composite; preserved in the source archive

**Figure 2.** Experimental design and registry-remediated composite. (a) Eighty tasks were crossed with seven conditions and three repetitions (1680 assignments) in 240 randomized task-repetition blocks. Expected values and tolerances were withheld from planner and runtime-validator interfaces; comparative estimates equally weighted 16 workflow clusters. (b) After incomplete unit metadata or prepared-evidence mappings were identified in ten tasks, their complete 21-assignment blocks were rerun. The composite retained 1470 unaffected original records and substituted 210 remediation records, selecting 1680 assignments from 1890 formal executions.

### 3.3. Flood Cases and Data Roles

The two cases were selected to represent distinct operational information settings in flood-GIS work (Table 3; Figure 3). Zhengzhou is a data-to-product case: pre- and post-event SAR observations are transformed into candidate flood, exposure, risk, and report products. Forlì is a product-to-decision-support case: a CEMS EMSR664 observed flood extent is already available at task start and is used to generate downstream exposure, risk, and reporting products while preserving its provenance. The Forlì case is therefore not an independent test of SAR flood extraction. Instead, it tests whether the evaluation frame-

work remains interpretable when a workflow begins from a prepared authoritative flood product rather than raw observations.

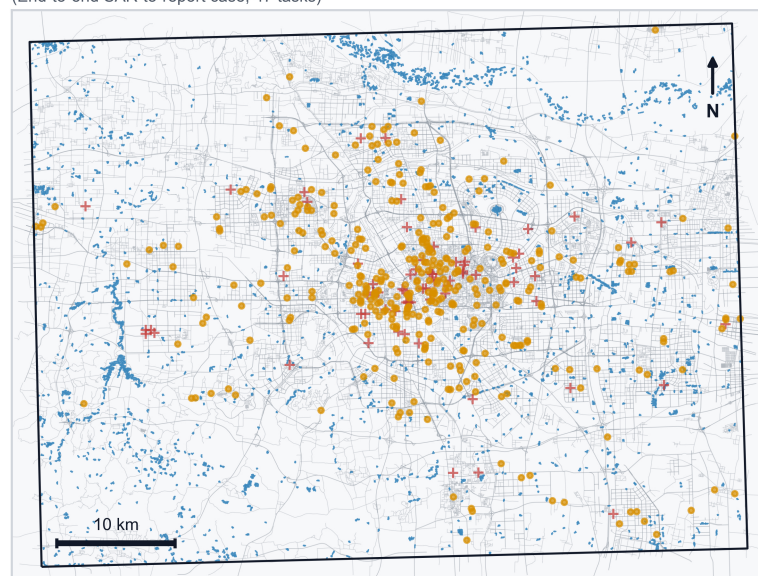
**Table 3.** Case scope and operational data.

| Case                                         | AOI and Event                                                                              | Operational Data Role                                                                                                                                                                                                | Evaluated Chain;<br>Excluded Inference                                                       |
|----------------------------------------------|--------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| Zhengzhou,<br>China, July 2021               | Main urban AOI,<br>113.30–113.95° E and<br>34.55–34.95° N; event<br>window 17–22 July 2021 | Pre/post Sentinel-1 VV/VH, GPM<br>IMERG V07, CHIRPS, Shuttle<br>Radar Topography Mission<br>(SRTM) digital elevation model<br>(DEM), JRC Global Surface Water<br>and OpenStreetMap<br>roads/facilities [38,40,43–46] | SAR → exposure → risk<br>→ report; workflow<br>conformance, not<br>flood-map accuracy.       |
| Forlì,<br>Emilia-Romagna,<br>Italy, May 2023 | AOI, 11.97–12.10° E and<br>44.17–44.28° N; event<br>window 16–24 May 2023                  | CEMS EMSR664 observed flood<br>extent as a reference-conditioned<br>operational input; OpenStreetMap<br>roads and facilities [46,47]                                                                                 | Prepared EMSR664 extent<br>→ exposure → risk →<br>report; not independent<br>SAR extraction. |

Abbreviations: AOI, area of interest; CEMS, Copernicus Emergency Management Service; CHIRPS, Climate Hazards Group InfraRed Precipitation with Station data; DEM, digital elevation model; GPM, Global Precipitation Measurement; IMERG, Integrated Multi-satellite Retrievals for GPM; JRC, Joint Research Centre; SAR, synthetic aperture radar; SRTM, Shuttle Radar Topography Mission; VH, vertical-transmit/horizontal-receive polarization; VV, vertical-transmit/vertical-receive polarization.

**a Zhengzhou, China · July 2021**

(End-to-end SAR-to-report case; 47 tasks)



**b Forlì, Italy · May 2023**

(Reference-conditioned downstream case; 33 tasks)



Map data © OpenStreetMap contributors; Zhengzhou candidate extent: Copernicus Sentinel-1 data (ESA); Forlì extent: Copernicus EMSR664.

**Figure 3.** Study areas and case roles. (a) Zhengzhou contributes 47 tasks spanning a synthetic aperture radar (SAR)-to-report workflow. (b) Forlì contributes 33 downstream tasks based on a prepared Copernicus Emergency Management Service (CEMS) observed-event flood extent. Black outlines indicate each analysis area of interest (AOI). Data sources: Sentinel-1 [38], OpenStreetMap [46], and CEMS EMSR664 [47]; locator outlines are from Natural Earth.

The Zhengzhou pre-event Sentinel-1 window was 20 June–10 July 2021 and the post-event window was 21 July–5 August 2021. Rainfall rasters were accumulated over the

half-open interval 17–22 July 2021, comprising observations from 17 July up to, but not including, 22 July. The Forlì flood extent and summary were registered as pre-existing operational inputs derived from the Copernicus Emergency Management Service (CEMS) EMSR664 observed-event product. The original CEMS layer was retained as reference-only provenance and was not presented as an output of GeoFloodAgent. Source identifiers, acquisition windows and license information were recorded in the case manifests; hashes for the 33 local files used at execution were recorded in the run-level case-data manifest.

The CEMS source uniform resource identifier (URI) and embedded layer name were retained as provenance metadata. The release archive does not treat the prepared products as independently reproducible SAR outputs. Because the Overpass exports lacked historical snapshot dates, the OpenStreetMap layers are treated as frozen contemporary infrastructure inventories rather than reconstructions of the road and facility networks at the 2021 or 2023 event dates.

The two cases test the same evaluation proposition under different starting conditions. In Zhengzhou, a contract-aligned result requires the requested products to be generated from registered case inputs and linked to the reported claim. In Forlì, a contract-aligned result may rely on the prepared observed extent when it satisfies the declared evidence requirement. Thus, Figure 3 visualizes study inputs and data roles rather than providing an independent accuracy validation of either flood extent. In neither case do the present results compare a derived flood layer with an external reference dataset for hydrological validation.

### 3.4. Flood-Task Benchmark Construction

We constructed 80 template-structured Chinese-language tasks from the two case manifests, deterministic tool outputs, and seven analysis categories. The benchmark represents designed analytic requirements rather than a sample of incident-command requests, analyst histories, or emergency standard operating procedures. The frozen registry records case, category, task form, nominal difficulty label, and gold-plan length from one to seven steps; Table 4 gives the exact counts. The nominal difficulty label is an author-assigned descriptor, not a validated measure of real-world flood complexity.

**Table 4.** Frozen benchmark composition.

| Dimension                     | Frozen Levels and Task Counts                                                                                                                |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| Case                          | Zhengzhou: 47; Forlì: 33                                                                                                                     |
| Category                      | case context: 10; flood mapping: 12; rainfall/terrain: 10; exposure analysis: 14; risk assessment: 14; report generation: 10; robustness: 10 |
| Difficulty                    | basic: 29; intermediate: 28; advanced: 23                                                                                                    |
| Gold trajectory length        | 1: 24; 2: 25; 3: 14; 4: 12; 7: 5                                                                                                             |
| Contract form                 | artifact-bearing: 70; capability-boundary robustness: 10                                                                                     |
| Paired-cluster reporting unit | 16 workflow clusters defined by case, task category, and gold tool sequence                                                                  |

Each task specifies a prompt, a minimal gold plan, required artifacts, permissible evidence artifacts, scoring weights, and one registered claim or action. Seventy tasks require at least one artifact; the ten robustness tasks are scored through a boundary action. The endpoint is intentionally narrow. Numeric expected values use an absolute tolerance of 0.01 when represented as floats, while integer, string, and Boolean claims require equal-

ity. The tolerance accommodates deterministic serialization of the reference outputs. It does not define scientific accuracy. For each of the ten report tasks, evaluation covers one registered metric; other qualifiers, denominators, statements about spatial scope, and unrestricted prose are not scored.

Boundary-action contracts followed a study-specific policy defined before execution. A clarification action was required when necessary case information or parameters were missing or conflicting. A refusal action was required for unsupported predictions, unjustified levels of spatial or temporal detail, unwarranted certainty claims, or autonomous high-stakes decisions. Each boundary action was terminal and did not invoke a GIS tool. Exact action matching therefore assessed agreement with the study policy; it did not assess the operational appropriateness or cost of clarification, refusal, or omission.

Reference artifacts were produced with the deterministic toolchain and used to construct candidate contracts. Before execution, two automated reviews applied the same seven-criterion rubric to all tasks in separate contexts. An author with surveying, mapping, and remote-sensing qualifications then examined a stratified sample of 16 tasks; all sampled contracts passed. This procedure served as internal quality control. The full registry was not independently reviewed by GIS specialists, and the author sample included only two capability-boundary tasks, both with the same refusal label. The task and trajectory files were hash-verified before execution.

To describe heterogeneity within the frozen benchmark, outcomes are additionally stratified by case, task form, nominal difficulty label, and gold-plan length. These descriptors were fixed in the registry before outcome summaries were calculated. Because task form, category, plan length, and nominal difficulty are partly co-determined in this designed benchmark, the stratified analysis is descriptive and does not estimate an independent causal effect of task complexity on agent performance. For transparency, we also tabulated the Zhengzhou/Forlì and category composition of each nominal-difficulty and gold-plan-length stratum from the frozen registry.

### 3.5. Experimental Conditions

We evaluated seven conditions (Table 5), six of which used the same hosted planner configuration. The primary exploratory comparison paired Full with the same-endpoint Tool Agent. Four component-withholding conditions probed the coupled implementation; as component interactions were not crossed factorially, these comparisons are descriptive ablations. Fixed Workflow served as a policy baseline and used a frozen, manually authored geoprocessing workflow in place of model-generated plans.

**Table 5.** Frozen condition matrix. LLM denotes large language model.

| Condition                | LLM Planner | Field-Level Schema Metadata | Evidence Ledger | Runtime Validator | Bounded Repair       |
|--------------------------|-------------|-----------------------------|-----------------|-------------------|----------------------|
| GeoFloodAgent (Full)     | Yes         | Yes                         | Yes             | Yes               | Yes                  |
| Same-endpoint Tool Agent | Yes         | Yes                         | No              | No                | No                   |
| Schema Metadata Withheld | Yes         | No                          | Yes             | Yes               | Yes                  |
| No Evidence              | Yes         | Yes                         | No              | Yes               | Yes                  |
| No Validator             | Yes         | Yes                         | Yes             | No                | Pre-execution only * |
| No Repair                | Yes         | Yes                         | Yes             | Yes               | No                   |
| Fixed Workflow           | No          | Not applicable              | No              | No                | No                   |

\* In the No Validator condition, plan-level checks remained active before execution. They could reject an invalid plan, but post-execution repair could not be triggered because runtime validation was disabled.

Tool Agent receives the same prompt, case context, layer identifiers, tool catalog, and field-level schemas as Full but omits the ledger, runtime validator, and repair. Their contrast is therefore system-level rather than an estimate of any single component's effect. The component-withholding configurations are also coupled. For example, No Evidence removes both ledger construction and ledger-dependent validation while retaining artifact, raster, geometry, and archive-profile checks.

Schema Metadata Withheld retains the case, layer, tool, dependency, artifact, and boundary-action information but removes field-level plan and argument schemas from the prompt. Provider JSON-object mode and the common strict parser remain active, so this condition tests planner–parser interoperability rather than unconstrained schema-free planning. Fixed Workflow does not call an LLM or inspect task gold plans; it executes every applicable stage in a frozen case-adaptive pipeline and represents a complete case-adaptive workflow rather than task-minimal planning.

### 3.6. Model Configuration and Randomized Interleaving

All six model-backed conditions used the official DeepSeek API (<https://api.deepseek.com/chat/completions> (accessed on 24 September 2026)) with the provider-advertised deepseek-v4-flash model, temperature 0, a 4096-token output limit, a 120 s timeout, and thinking disabled. The original execution campaign ran from 14 to 16 July 2026 UTC, and the ten-task remediation campaign ran on 17 July 2026 UTC. The requested and returned model labels were recorded. Because the hosted endpoint did not expose fixed weights, routing, or provider-side update history, this study evaluates the named endpoint configuration during the recorded execution windows rather than a stable model architecture. Local deployment was not used because it would introduce a different model and serving configuration, rather than control the behavior of the tested hosted endpoint. Prospective replication with version-pinned locally deployed models is therefore required to assess generality beyond this endpoint.

Each task was assigned to all seven conditions in three repetitions. The resulting 240 task-repetition blocks contained every condition once, with block and within-block order fixed before execution (Figure 2a). Runs used separate temporary workspaces with caching disabled. Before a workspace was removed, a condition-agnostic snapshot captured the content, checksum, and spatial structure required for offline assessment.

After execution, an archive audit identified incomplete unit metadata or prepared-evidence mappings in ten tasks. For each affected task, we reran the complete 21-assignment block (seven conditions  $\times$  three repetitions) and replaced the corresponding original block in the analytic dataset. The resulting composite contains 1470 original and 210 replacement records. Complete-block replacement avoided outcome-selective rerunning. However, because the affected tasks were identified post hoc and the replacement blocks were executed later against a hosted endpoint, this design cannot separate registry correction from temporal endpoint variation. We therefore report the original archive as a sensitivity analysis. This limitation is particularly important for model-agent studies: registry correction and endpoint time cannot be separated in the present archive, and the results should not be interpreted as invariant to future provider updates.

### 3.7. Interface-Level Target Withholding and Contract Assessment

Frozen reference values, numerical tolerances, and gold plans were withheld from planner prompts and runtime-validator interfaces. After each run, the system serialized the required artifacts and observed boundary action without exposing those target values or tolerances; observations that contained them were rejected. A separate archive evaluator then resolved each registered claim from an admissible artifact and compared it with the

frozen reference value. For JSON products, it followed declared metric paths; for reports, it used prespecified metric-specific extraction patterns. Diagnostic gold-plan scores did not affect runtime acceptance, validation feedback, repair, or either primary endpoint.

Runs used run-specific workspaces, but the design implements interface-level target withholding rather than process-level or operating-system security isolation. Offline contract alignment denotes agreement with a frozen, study-authored product or action specification. It is an evaluation construct, not a test of unrestricted report content, hydrological accuracy, cybersecurity, or operational decision quality. Because the tools, registry, reference artifacts, and evaluator share an implementation lineage, target withholding cannot eliminate common specification errors.

Runtime disposition and agreement with the registered task contract form four mutually exclusive outcomes:

- contract-aligned acceptance: accepted and aligned with the offline contract;
- accepted contract mismatch: accepted but not aligned with the offline contract;
- rejected contract mismatch: rejected and not aligned with the offline contract; and
- runtime rejection despite contract alignment: rejected although aligned with the offline contract.

A record is accepted only when execution succeeds and, if the condition includes a validator, final validation passes. For conditions without a validator, successful execution constitutes runtime acceptance.

### 3.8. Outcomes and Descriptive Analyses

The outcome analysis reports the four mutually exclusive runtime-disposition  $\times$  contract-alignment cells and the task-form composition of accepted mismatches and boundary actions. The two comparative endpoints are accepted-contract-mismatch incidence and contract-aligned-acceptance rate, each calculated over all assigned runs. The former measures runs that were accepted at runtime but did not align with their registered task contracts, whereas the latter measures runs that were both accepted at runtime and contract-aligned. Using all assigned runs provides a common denominator across conditions. Reporting the contract-aligned-acceptance rate prevents a zero mismatch rate caused by universal rejection from being interpreted as safety. We also report mismatches among accepted runs as a descriptive conditional rate because acceptance itself varies by condition.

Additional descriptive analyses summarized repair, latency, boundary actions, registry-remediation sensitivity, and validator exposure. Ordinary connection, timeout, overload, and rate-limit failures remained assigned-run outcomes. Ancillary analyses report task-structure stratification and deterministic parameter sensitivity separately from the primary system comparison. Neither ancillary analysis changes the benchmark denominator, the contract-alignment definitions, or the Full-Tool Agent estimand.

### 3.9. Statistical and Ancillary Descriptive Analyses

The Full-Tool Agent comparison uses paired-cluster comparisons on a finite benchmark. Tests and resampling intervals are exploratory and are not used for confirmatory causal or population-level claims.

For each condition and task, the three binary repetition outcomes are first averaged. The Full-Tool Agent estimand is the equal-weight mean paired difference across 16 workflow clusters defined from the frozen case, task-category, and gold-tool-sequence fields. Within each cluster, task-level repetition means are averaged; cluster means then receive equal weight, so a large family of near-duplicate tasks cannot dominate the summary. The

estimand targets this finite 16-cluster benchmark under equal cluster weighting; it does not target the population of flood events or GIS requests.

Let  $Y_{t,c,r}$  denote a binary outcome for task  $t$ , condition  $c$  and repetition  $r$ , and let  $\mathcal{C}_k$  denote the tasks in cluster  $k$ . For Full (F) versus Tool Agent (T), the estimator is

$$\bar{Y}_{t,c} = \frac{1}{3} \sum_{r=1}^3 Y_{t,c,r} \quad (1)$$

$$\hat{\Delta}_{F-T} = \frac{1}{K} \sum_{k=1}^K \left[ \frac{1}{n_k} \sum_{t \in \mathcal{C}_k} (\bar{Y}_{t,F} - \bar{Y}_{t,T}) \right] \quad (2)$$

where  $K = 16$  and  $n_k = |\mathcal{C}_k|$ . The same estimator is used for each binary endpoint and Full-comparator contrast.

We obtained two-sided sign-flip  $p$  values by enumerating all sign assignments of the 16 cluster differences, retaining zeros in the estimand denominator. Exactness requires joint invariance to independent sign changes, an assumption not guaranteed because clusters share events, tools, and task templates; the  $p$  values are therefore exploratory. We estimated uncertainty from 5000 with-replacement resamples of the 16 paired cluster differences (seed 20260714) and report the equal-tailed 2.5th and 97.5th percentiles [48]. These empirical intervals quantify sensitivity to the observed cluster composition rather than population coverage and are not obtained by inverting the sign-flip test.

The primary Full-Tool Agent analysis reports the mean percentage-point difference, a 95% empirical interval, and raw and Holm-adjusted  $p$  values, with Holm correction across the two endpoints [49]. Alternative-condition contrasts were adjusted across the five comparators within each endpoint; adjustment across all ten contrasts was treated as a multiplicity sensitivity analysis.

With 16 workflow clusters and no prospective power calculation, all inference was exploratory. Leave-one-cluster-out ranges summarized cluster influence rather than uncertainty. Fixed Workflow was interpreted as a policy comparison. Because the complete analysis procedure lacked an independently timestamped pre-execution record, non-significance was not interpreted as evidence of equivalence.

### 3.10. Post Hoc Contract-Implementation Diagnostics

To assess whether the validator implemented the registered contract predicates as intended, we conducted two post hoc offline diagnostics. A static compatibility check traversed all 70 artifact-bearing contracts against their reference artifacts and ledger-extraction paths. For non-JSON formats, the static compatibility check populated a synthetic ledger candidate from the registered expected value and separately checked the reference artifact for a matching textual representation. A mutation study then applied pre-defined defect families to eligible artifacts. Candidate-selection rules, mutation operators, expected issue codes, identity controls, exclusions, reporting rules, and source and code hashes were fixed before validator execution. The eligible corpus comprised eight required GeoJSON products, including six suitable for type-preserving invalid-geometry mutations; 14 evidence-source clusters, including 12 with units and ten with JSON sources; and five auxiliary GeoTIFFs in a formally unexposed raster branch. A secondary sweep covered 68 unique case-source-metric profiles.

Each mutation was paired with an applicable identity-rewrite control; a unique control could be reused across defect families when they shared the same artifact or source cluster. Incorrect-but-present CRS labels, coordinate swaps or translations, null geometries, and coherent source-ledger value changes were treated as scope probes because no registered predicate targeted those semantics. These probes were excluded from detection-rate denominators. Expected-code detection and control rejection were summarized by

defect family using descriptive 95% Wilson intervals [50]. The mutation diagnostic ran without model inference or network access and assessed implementation response to a finite constructed corpus; it did not estimate recall for unregistered spatial errors.

Before analysis, automated checks confirmed archive coverage, schedule balance, target-withheld observation-schema validation, per-run workspace handling, route provenance, file integrity, and the absence of duplicate or unbalanced cells. Smoke-test and sandbox-connection records were excluded.

## 4. Results

Unless noted, counts and paired estimates use the registry-remediated composite (1680 assignments). Figure 2b and Section 4.3 document its construction.

### 4.1. Runtime Disposition $\times$ Contract Alignment

Table 6 reports the four runtime-disposition  $\times$  contract-alignment cells for all conditions, whereas Figure 4 separates the task-form composition of accepted mismatches from the exploratory Full-Tool Agent contrast.

**Table 6.** Runtime disposition and the four mutually exclusive contract-alignment outcomes. Percentages use all 240 assigned runs per condition. The repair column gives runs with at least one repair attempt, not successful recoveries.

| Condition                      | Accepted,<br><i>n</i> (%) | Contract-<br>Aligned<br>Acceptance, <i>n</i><br>(%) | Accepted<br>Contract<br>Mismatch, <i>n</i><br>(%) | Rejected<br>Contract<br>Mismatch, <i>n</i><br>(%) | Runtime<br>Rejection<br>Despite<br>Contract<br>Alignment, <i>n</i><br>(%) | Repair<br>Activated, <i>n</i><br>(%) |
|--------------------------------|---------------------------|-----------------------------------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------------------------------|--------------------------------------|
| GeoFloodAgent<br>(Full)        | 228 (95.0)                | 212 (88.3)                                          | 16 (6.7)                                          | 9 (3.8)                                           | 3 (1.3)                                                                   | 14 (5.8)                             |
| Same-endpoint<br>Tool Agent    | 228 (95.0)                | 208 (86.7)                                          | 20 (8.3)                                          | 9 (3.8)                                           | 3 (1.3)                                                                   | 0 (0.0)                              |
| Schema<br>Metadata<br>Withheld | 0 (0.0)                   | 0 (0.0)                                             | 0 (0.0)                                           | 240 (100.0)                                       | 0 (0.0)                                                                   | 240 (100.0)                          |
| No Evidence                    | 231 (96.3)                | 214 (89.2)                                          | 17 (7.1)                                          | 6 (2.5)                                           | 3 (1.3)                                                                   | 18 (7.5)                             |
| No Validator                   | 233 (97.1)                | 210 (87.5)                                          | 23 (9.6)                                          | 3 (1.3)                                           | 4 (1.7)                                                                   | 10 (4.2)                             |
| No Repair                      | 223 (92.9)                | 210 (87.5)                                          | 13 (5.4)                                          | 14 (5.8)                                          | 3 (1.3)                                                                   | 0 (0.0)                              |
| Fixed Workflow                 | 240 (100.0)               | 210 (87.5)                                          | 30 (12.5)                                         | 0 (0.0)                                           | 0 (0.0)                                                                   | 0 (0.0)                              |

Percentages in this table are unweighted proportions over 240 assigned runs per condition. Comparative estimates in Sections 4.2–4.4 use equal weighting across the 16 workflow clusters.

Table 6 shows why acceptance alone is insufficient. Schema Metadata Withheld records no accepted mismatches because it accepts no run, not because it produces aligned outputs. Full and Tool Agent each produced three runtime rejections despite contract alignment, so runtime disposition and registered-product evidence can disagree in either direction. Among accepted runs, the descriptive conditional mismatch rates were 7.0% for Full and 8.8% for Tool Agent.

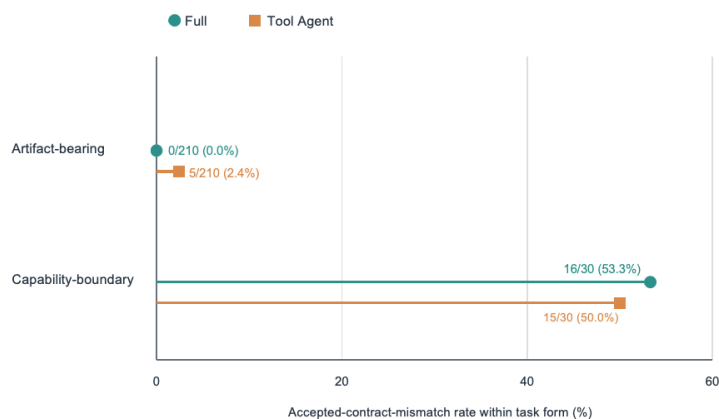
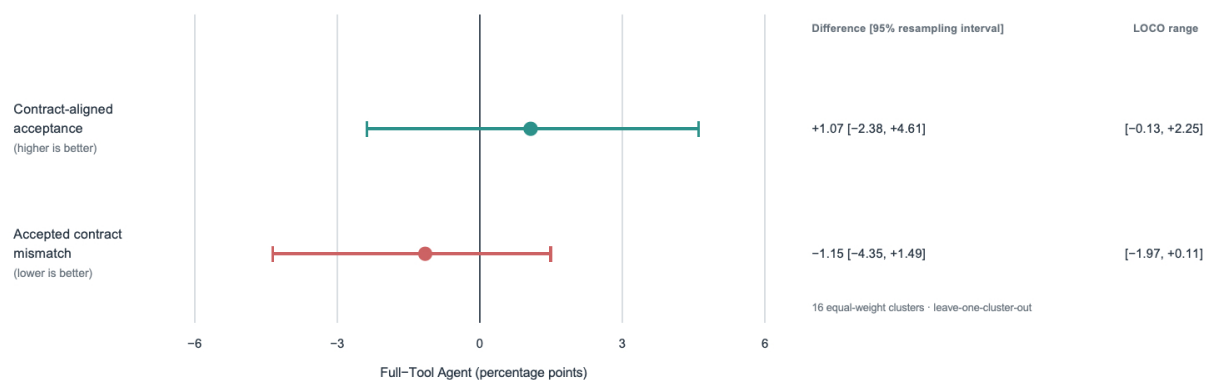
**a Accepted-contract-mismatch rates by task form****b Exploratory Full–Tool Agent differences**

**Figure 4.** Task-form composition and exploratory Full–Tool Agent comparisons. **(a)** Accepted-contract-mismatch rates by task form; Full’s 16 boundary mismatches comprise six incorrect refusals and ten omitted actions. **(b)** Equal-cluster-weighted Full-minus-Tool-Agent differences across 16 workflow clusters. Points show estimates, bars show 95% empirical cluster-resampling intervals, and brackets show leave-one-cluster-out influence ranges rather than confidence intervals.

#### 4.2. Descriptive Task-Structure Stratification

The outcome panel of Table 7 shows no monotonic association between the author-assigned nominal-difficulty labels and contract-aligned acceptance. For Full, contract-aligned acceptance was 90.8%, 82.1%, and 92.8% across basic, intermediate, and advanced strata, respectively; the corresponding Tool Agent rates were 87.4%, 82.1%, and 91.3%. The composition panel makes the overlap among task descriptors explicit: all five 5–7-step tasks are Zhengzhou report-generation tasks, whereas all ten capability-boundary tasks occur in the 1–2-step group. The stratification therefore describes performance within the frozen registry and helps delimit the evidenced capability boundary; it is not a balanced test of an independent complexity effect.

Figure 5 grounds the two evaluation axes in representative Full-configuration tasks. Zhengzhou shows concordant runtime acceptance and offline alignment, whereas Forli shows that a rejected execution can remain aligned when valid prepared evidence already satisfies the registered contract. This contrast demonstrates why runtime disposition and contract alignment must be reported separately.

**Table 7.** Descriptive contract-alignment outcomes by nominal difficulty and gold-plan length, with frozen-benchmark composition.

| Panel A. Descriptive contract-alignment outcomes |              |       |              |              |              |              |
|--------------------------------------------------|--------------|-------|--------------|--------------|--------------|--------------|
| Descriptor                                       | Level        | Tasks | Full CAA (%) | Full ACM (%) | Tool CAA (%) | Tool ACM (%) |
| Difficulty                                       | Basic        | 29    | 90.8         | 0.0          | 87.4         | 5.7          |
| Difficulty                                       | Intermediate | 28    | 82.1         | 14.3         | 82.1         | 14.3         |
| Difficulty                                       | Advanced     | 23    | 92.8         | 5.8          | 91.3         | 4.3          |
| Plan length                                      | 1–2 steps    | 49    | 83.0         | 10.9         | 82.3         | 12.2         |
| Plan length                                      | 3–4 steps    | 26    | 96.2         | 0.0          | 92.3         | 2.6          |
| Plan length                                      | 5–7 steps    | 5     | 100.0        | 0.0          | 100.0        | 0.0          |

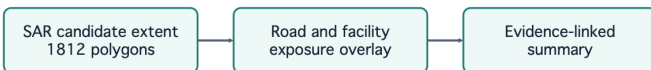
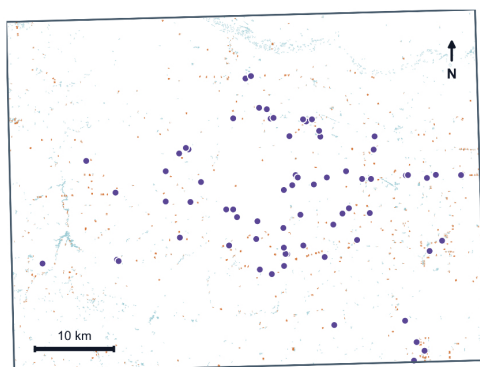
| Panel B. Frozen-benchmark composition |              |       |                 |                                       |
|---------------------------------------|--------------|-------|-----------------|---------------------------------------|
| Descriptor                            | Level        | Tasks | Zhengzhou/Forlì | Category composition                  |
| Difficulty                            | Basic        | 29    | 17/12           | CC4, FM5, RT4, EA6, RA6, RG4, RB0     |
| Difficulty                            | Intermediate | 28    | 17/11           | CC4, FM4, RT4, EA4, RA4, RG4, RB4     |
| Difficulty                            | Advanced     | 23    | 13/10           | CC2, FM3, RT2, EA4, RA4, RG2, RB6     |
| Plan length                           | 1–2 steps    | 49    | 28/21           | CC10, FM12, RT10, EA7, RA0, RG0, RB10 |
| Plan length                           | 3–4 steps    | 26    | 14/12           | CC0, FM0, RT0, EA7, RA14, RG5, RB0    |
| Plan length                           | 5–7 steps    | 5     | 5/0             | CC0, FM0, RT0, EA0, RA0, RG5, RB0     |

CAA, contract-aligned acceptance; ACM, accepted contract mismatch. Panel A values are percentages of all assigned runs within the indicated stratum. Panel B gives task counts from the frozen registry. CC, case context; FM, flood mapping; RT, rainfall–terrain; EA, exposure analysis; RA, risk assessment; RG, report generation; RB, robustness/capability-boundary. The composition differs across strata; Panel A therefore describes outcome patterns in this designed benchmark rather than an independent effect of any single complexity attribute.

#### a Zhengzhou: runtime accepted; contract aligned

SAR-derived candidate extent and exposure products

■ Flood candidate    — Flood-intersecting segments  
● Facilities within 200 m    — Analysis area



Runtime  
**ACCEPTED (3/3)**

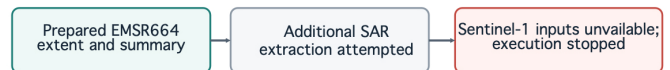
Contract alignment  
**ALIGNED**

Contract-checked measure  
**Flood-road overlap  
102.975 km**

#### b Forlì: runtime rejected; contract aligned

Prepared EMSR664 extent available at task start

■ Prepared extent    — Road context    — Analysis area



Runtime  
**REJECTED (3/3)**

Contract alignment  
**ALIGNED**

Contract-checked measure  
**Prepared flood  
polygons: 37**

**Figure 5.** Representative Full-condition outcomes. (a) Zhengzhou used a synthetic aperture radar (SAR)-derived candidate extent; all three runs were accepted and contract-aligned. (b) Forlì used a prepared Copernicus Emergency Management Service (CEMS) EMSR664 extent; all three runs were rejected when unavailable Sentinel-1 inputs stopped additional extraction, but remained contract-aligned because the prepared extent satisfied the registered requirement. Dark outlines mark the analysis area of interest (AOI). Data sources: Sentinel-1 [38], OpenStreetMap [46], and CEMS EMSR664 [47].

The 16 Full boundary mismatches comprised six refusals where clarification was required and ten omitted actions. One additional omitted refusal was rejected at runtime, while all six assignments involving autonomous high-stakes requests matched the registered refusal action. Because the action labels encode a study-authored policy rather than independently adjudicated operator decisions, these counts measure policy alignment rather than operational correctness. They are reported without a severity-weighted score because the costs of delay, omission, and unsafe continuation were not elicited from operators.

#### 4.3. Exploratory Full–Tool Agent Rates

Figure 4b and Table 8 report cluster-weighted Full-minus-Tool-Agent contrasts on the two exploratory comparative endpoints. Only three of the 16 workflow clusters had a nonzero accepted-mismatch difference, and five had a nonzero contract-aligned-acceptance difference.

**Table 8.** Exploratory Full–Tool Agent comparison.

| Comparative Endpoint                                   | Full, n/N (%)  | Tool Agent, n/N (%) | Paired Difference, pp<br>(95% Empirical Interval) | Raw <i>p</i> | Holm <i>p</i> |
|--------------------------------------------------------|----------------|---------------------|---------------------------------------------------|--------------|---------------|
| Accepted-contract-mismatch incidence (lower is better) | 16/240 (6.7)   | 20/240 (8.3)        | −1.15 (−4.35, 1.49)                               | 0.750        | 1.000         |
| Contract-aligned-acceptance rate (higher is better)    | 212/240 (88.3) | 208/240 (86.7)      | 1.07 (−2.38, 4.61)                                | 0.625        | 1.000         |

Event counts and rates use all 240 assigned runs in each condition. Paired differences are Full minus Tool Agent, averaged with equal weight over the 16 workflow clusters defined from frozen task fields. Empirical intervals use paired cluster resampling; *p* values are from two-sided enumerated sign-flip tests, exact only under the joint sign-change invariance stated in Section 3.9, with the stated multiplicity adjustment. Both are interpreted exploratorily, not as confirmatory population inference. Abbreviation: pp, percentage points.

Cluster-weighted differences were −1.15 percentage points for accepted contract mismatch and +1.07 percentage points for contract-aligned acceptance. Both 95% intervals included benefit and harm (Holm-adjusted *p* = 1.000). On this benchmark, the estimates do not demonstrate an advantage for Full over Tool Agent, nor do they establish practical equivalence. The comparison shows why the classes in Section 4.1 remain necessary even when headline rates are similar: an execution-success metric would merge wrong refusals, omitted actions, and artifact-bearing mismatches while hiding the rejection floor in Table 6.

The leave-one-cluster-out analysis reached the same conclusion. The accepted-mismatch difference ranged from −1.97 to +0.11 percentage points and the contract-aligned-acceptance difference from −0.13 to +2.25 percentage points after omitting one workflow family at a time (Figure 4b).

#### 4.4. Sensitivity to Registry Remediation

Figure 2b summarizes the construction of the analysis dataset. It combines 1470 unaffected assignments from the original schedule with 210 complete-block remediation assignments for ten tasks whose unit or prepared-evidence mappings were incomplete (1680 of 1890 formal executions). The accepted-mismatch contrast was unchanged at −1.15 percentage points because neither Full nor Tool Agent produced an accepted mismatch in the affected task blocks. By contrast, the contract-aligned-acceptance difference changed from −7.68 percentage points in the original archive to +1.07 percentage points in the remediated composite.

The observed change should not be interpreted as the causal effect of registry remediation. The replacement block was executed later against a hosted endpoint whose fixed weights and routing versions were unavailable, and several other conditions also changed on at least one affected-block outcome. Registry correction and execution time are therefore inseparable in this dataset.

#### 4.5. Comparisons with Coupled Configurations

Schema Metadata Withheld rejected all 240 assignments. Its zero accepted-mismatch rate therefore reflected universal non-acceptance rather than improved alignment. With equal weight assigned to the 16 workflow clusters, the Full–Schema Metadata Withheld difference in contract-aligned acceptance was 79.28 percentage points. This cluster-weighted estimand differs from the unweighted 88.3-percentage-point contrast obtained from the assignment-level counts in Table 6. This was the only contrast that remained below 0.05 under both the endpoint-specific and joint ten-contrast Holm corrections (adjusted  $p = 0.00061$  and  $0.00122$ , respectively); all other adjusted comparisons were at least 0.625.

Contract-aligned-acceptance rates for Full, No Evidence, No Validator, No Repair, and Fixed Workflow ranged from 87.5% to 89.2%, although No Validator and Fixed Workflow produced more accepted mismatches than Full. Because these conditions altered coupled system configurations rather than individual components in a factorial design, the contrasts characterize complete configurations and do not identify the independent effect of evidence capture, validation, or repair.

#### 4.6. Repair Behavior and Validator Scope

Full invoked repair in 14 runs (5.8%), generating 24 attempts; six runs were recovered and eight remained unrecovered. No Evidence recovered 12 of its 18 repair-activated runs. In No Validator, pre-execution plan repair recovered seven of ten repair-activated runs; post-execution repair was unavailable because runtime validation was disabled. Schema Metadata Withheld exhausted the two-attempt allowance in every run without recovery, indicating that repeated prompting did not overcome the missing field-level schema under the shared parser.

On unmodified Full runs, none of the geometry, archive-profile, unit, source-identifier/path, or source-value checks failed, and only two metric-support checks failed. These observations describe validator exposure under normal outputs rather than detection performance for corrupted inputs. No formal task required a GeoTIFF artifact, so the raster branch had no normal-output exposure.

The static compatibility audit found no inconsistencies among the 70 artifact-bearing contracts under this reference-artifact and ledger-path construction; the ten action contracts were not applicable to ledger linkage. Because the audit did not replay formal observations or independently recompute registered values, this result indicates internal compatibility rather than independent validity.

In the post hoc mutation study, the expected issue code appeared in all registered primary cases: 8/8 archive-profile removals, 8/8 empty products, 6/6 type-preserving invalid geometries, 14/14 cases in each of four evidence-source families, 12/12 unit changes, and 10/10 cases in each of the missing-metric and changed-source-value families. None of the 22 unique primary identity controls were rejected. The auxiliary raster branch also produced the expected code for both missing-CRS and all-NoData mutations in 5/5 cases, with no rejection among five controls. The secondary profile sweep produced the expected code in all 411 constructed cases and rejected none of the 68 controls, although these rows reused 14 source clusters. By contrast, none of the 47 scope probes targeting unregistered spatial semantics triggered an issue. The descriptive 95% Wilson lower bounds for the

registered primary families ranged from 0.610 to 0.785. These findings document implementation response to the registered predicates but do not estimate recall for spatial errors in general.

Median latency was 4.7 s for Full, 4.0 s for Tool Agent, and 260.3 s for Fixed Workflow. Within Full, median latency was 8.4 s in the 14 repair-activated runs and 4.1 s in the 226 runs without repair; this descriptive contrast is confounded by task composition.

#### 4.7. Deterministic Parameter Sensitivity Analysis

Table 9 reports the completed frozen-input deterministic sensitivity analysis. In Zhengzhou, lowering the minimum SAR object size from the 80-pixel baseline to 40 pixels increased candidate flood extent by 105.6%, produced 8277 polygons, and increased flood-intersecting road length and facilities within the 200 m buffer by 244% and 218%, respectively. Increasing the threshold to 120 pixels reduced candidate extent by 30.9%, yielded 777 polygons, and reduced the corresponding road-length and facility summaries by 59% and 63%. Thus, the reported Zhengzhou exposure summaries are materially conditional on the pre-frozen denoising threshold.

For facility proximity, Zhengzhou returned 34, 73, and 102 facilities at 100, 200, and 300 m, respectively. Relative to the 200 m baseline, the 100 m setting changed 60 of 1248 risk-grid scores and 27 risk classes (2.16%); its high/medium/low class counts were 2/49/1197 rather than 3/30/1215. In Forlì, the same buffer settings returned 1, 1, and 5 facilities, and the 100 m setting changed neither risk-grid scores nor classes among 24 cells. These findings show that buffer choice can affect local exposure summaries, while the observed effect depends on case composition.

For vector simplification, Zhengzhou's 0 m tolerance changed flood-intersecting road length by 0.3% relative to the 10 m baseline; the 20 m setting reproducibly failed during geometric union, so no downstream summary was produced. In Forlì, the 0 m and 20 m settings changed road length by +1.4% and −0.8%, respectively, relative to the 10 m baseline. The analysis is an implementation-specific robustness diagnostic: it neither recomputes the primary LLM-agent endpoints nor validates flood-map accuracy or a universal optimal parameter set.

**Table 9.** One-factor-at-a-time deterministic GIS sensitivity analysis under frozen inputs.

| Observed Result                                                                                                                                   | Baseline and Tested Levels                                                                            | Case      | Parameter                |
|---------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|-----------|--------------------------|
| 40 px: area +105.6%; 8277 polygons; road length +244%; facilities +218%.<br>120 px: area −30.9%; 777 polygons; road length −59%; facilities −63%. | 80 px: 26.195 km <sup>2</sup> ; 1812 polygons; 102.975 km roads; 73 facilities. Levels: 40/80/120 px. | Zhengzhou | Minimum SAR object size  |
| Facilities: 34/73/102. At 100 m: 60/1248 scores changed; 27/1248 (2.16%) classes changed; high/medium/low = 2/49/1197.                            | 200 m: 73 facilities; 1248 grids (3/30/1215 high/medium/low). Levels: 100/200/300 m.                  | Zhengzhou | Facility buffer          |
| Facilities: 1/1/5. At 100 m, 0/24 grid scores or classes changed.                                                                                 | 200 m: 1 facility; 24 grids. Levels: 100/200/300 m.                                                   | Forlì     | Facility buffer          |
| 0 m: road length +0.3%. 20 m: reproducible geometric-union tool failure; no downstream summary.                                                   | 10 m: 102.975 km roads. Levels: 0/10/20 m.                                                            | Zhengzhou | Simplification tolerance |
| 0 m: +1.4%; 20 m: −0.8%.                                                                                                                          | 10 m: 12.330 km roads. Levels: 0/10/20 m.                                                             | Forlì     | Simplification tolerance |

Forlì did not undergo the minimum SAR object-size test because it began with a prepared EMSR664 observed extent. Percentages are relative to the case-specific frozen baseline. The Zhengzhou 20 m run did not complete because of a reproducible geometric-union tool failure. This analysis does not validate hydrological accuracy.

## 5. Discussion

### 5.1. Runtime Acceptance and Evidential Adequacy Are Distinct Outcomes

Joint reporting changed the interpretation of the observed failures. Universal rejection can eliminate accepted mismatches without producing aligned completions, whereas a run can be rejected even when adequate product evidence is available. Full and Tool Agent had similar, cluster-sensitive estimates. In this experiment, the four-cell outcome pattern is more informative than the small difference between their aggregate rates.

The boundary tasks extend this distinction beyond file production. Under the study-authored action policy, all 16 Full accepted mismatches involved incorrect or omitted clarification or refusal actions. The absence of accepted contract mismatches among Full's artifact-bearing assignments should be interpreted cautiously because the normal outputs contained few relevant defects.

### 5.2. Geospatial Product Contracts as an Intermediate Assurance Layer

The contracts govern specific flood, exposure, and risk products by linking each to required geometry, metadata, units, and supporting artifacts. This product-level view can reveal a missing risk surface or a road-length claim traced to the wrong exposure layer; generic execution provenance alone may not capture these problems. Product conformance thus occupies an intermediate position between successful geoprocessing and scientific or operational validity (Figure 1b).

This role is narrower than the spatial-quality frameworks described by ISO 19157-1 and OGC Simple Features [22,23]. The legacy top-level crs predicate further shows that conformance depends on the stated product profile: it supports an internal archive convention and should not be read as RFC 7946 compliance or CRS-correctness validation [41,42].

These checks were exercised during ordinary runs, but geometry, archive-profile, unit, source-path, and source-value defects were largely absent. The mutation corpus adds a direct implementation check for the registered predicates across the available artifacts and source clusters. The scope probes show where that coverage ends: a present CRS label may be wrong, coordinates may be swapped or displaced, a feature geometry may be null, and internally consistent lineage may still support a corrupted value. Deterministic geospatial libraries reduce arbitrary geometry mutation after correct tool execution, but input quality, coordinate semantics, cross-layer consistency, and scientific interpretation still require separate validation.

### 5.3. Schemas, Registries, and Coupled Controls

Schema withholding exposed an interface-compatibility floor. Without field-level schemas, the tested provider–prompt–parser combination did not produce a plan accepted by the shared parser. The resulting absence of accepted mismatches reflects failure to enter the executable workflow, not evidence that a schema alone makes the system safe.

The registry audit also exposes a measurement issue: units, permitted source artifacts, extraction rules, and reference values are components of both the runtime system and the measurement instrument. Incomplete registry metadata may therefore produce systematic rejection without any change in the underlying spatial evidence. Complete-block replacement avoided selecting favorable reruns, but registry correction and the later execution window were introduced together. A prospective replication with the corrected registry is required to estimate the contrasts without this temporal mixture.

Evidence capture, validation, and repair also operate as a coupled control loop. The single-component removals cannot reveal interactions such as Validator  $\times$  Repair, and the present contrasts should be read as observations of configured systems rather than estimates of component necessity. A full-factorial or staged intervention design would pro-

vide a stronger test of those mechanisms. In an operational study, that design should also include human review because an unsafe continuation, a wrong refusal, and a delayed clarification have different costs that cannot be inferred from contract alignment alone.

#### 5.4. Practical Implications and Benchmark Positioning

At a practical level, the proposed distinction can serve as an audit layer within, rather than in place of, conventional flood-GIS practice. After an analyst or an agent generates a candidate flood extent, an exposure summary, or a risk grid, the ledger and task contract can make it possible to check whether a reported length, area, count, or action is linked to a retained spatial product and declared source layers. This may improve auditability, hand-over, and human review in emergency GIS workflows. It does not authorize autonomous alerting, resource allocation, or hazard classification.

GeoFloodAgent is complementary rather than directly numerically comparable to GeoAgentBench and DORA. GeoAgentBench emphasizes dynamic task execution and spatial outputs in a broad tool sandbox [15], whereas DORA emphasizes heterogeneous disaster operations across events and geospatial tools [20]. The present benchmark contains two cases and a fixed seven-tool environment; its contribution is the post-run separation of runtime disposition from agreement with a task-specific product, claim, or action contract. Because task sources, tool inventories, model endpoints, and outcome definitions differ, a numerical ranking across these benchmarks would be invalid. A useful next step is to implement the present contract layer within larger, independently curated evaluation suites.

#### 5.5. Limitations of the Study

This study has several limitations related to its benchmark construction. The 80 tasks are synthetic and template-structured, and the contracts, reference artifacts, tools, and evaluator share a study implementation lineage. Only 16 tasks received an author audit, and the full registry was not independently reviewed by GIS specialists. Interface-level target withholding reduces direct exposure of reference targets but cannot remove common specification errors.

The two cases provide contrasting evidential settings but limited geographic and operational diversity. Zhengzhou evaluates a SAR-to-report workflow, whereas Forli begins from a prepared CEMS EMSR664 observed extent and is not an independent SAR-reproduction experiment. OpenStreetMap layers are frozen contemporary snapshots rather than event-date infrastructure inventories. The nominal difficulty labels and gold-plan lengths are task descriptors, not a validated measure of flood-case complexity; consequently, the added stratified summaries cannot establish a general complexity-performance relationship.

The deterministic parameter-sensitivity analysis is limited to selected post-processing defaults under frozen inputs. It shows that Zhengzhou summaries are materially conditional on the 80-pixel candidate-object threshold, that a 100 m rather than 200 m buffer changes a small proportion of risk-grid classes, and that a 20 m simplification tolerance can fail in the current workflow. These results are configuration-specific diagnostics; they do not calibrate the SAR change-detection procedure against an independent inundation reference, identify hydrologically optimal settings, or establish a universal transfer rule. Broader validation would require event-specific reference maps, uncertainty characterization, and testing across additional flood types and sensor conditions.

The primary comparisons use a hosted endpoint whose weights, routing, and update history are unavailable. In addition, registry remediation and the later execution window are inseparable in the composite archive. The reported intervals therefore describe the observed finite benchmark and should not be interpreted as stable model properties,

causal component effects, or evidence of equivalence. Prospective replication with version-pinned local and hosted models is needed.

Finally, product conformance and metric-artifact lineage are intermediate assurance properties, not operational validity. The study does not measure analyst workload, explanation quality, the consequences of false clarification or refusal, or the usefulness and authorization of outputs during a real incident. Independent contract review, historical requests, and operator-centered studies are necessary before the approach can support operational deployment.

Future work should use an externally reviewed multidimensional complexity framework that independently varies data-source count, spatial-operation count, dependency depth, provenance requirements, and boundary-action requirements.

## 6. Conclusions

Within this two-case, 80-task benchmark, runtime acceptance and offline contract alignment were not interchangeable. Universal non-acceptance could suppress accepted contract mismatches without producing aligned completions, and a run could be rejected despite retaining evidence that satisfied the registered contract. The Full and same-endpoint Tool Agent configurations did not show a stable comparative advantage. The principal empirical lesson is therefore that LLM-orchestrated flood-GIS workflows should report the four runtime-disposition  $\times$  contract-alignment outcomes rather than a single completion rate.

GeoFloodAgent contributes an intermediate audit layer that links selected flood, exposure, and risk products to declared structural requirements and metric lineage. It can support inspection of whether a reported claim is backed by a retained spatial product, but it does not independently establish flood-map accuracy, safety, or emergency-decision utility. Independent contracts, prospective version-controlled replications, additional flood settings, hydrological reference validation, and operator studies are required before the approach can be assessed for operational use.

**Author Contributions:** Conceptualization, Jin Liu; methodology, Hongyun Zhang; software, Ye Xuan; validation, Yahong Zhao; writing-review and editing, Fang Wang. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Fundamental Research Funds for the Central Universities under Grant 3142024048 and Grant 3142024045, and the National Key Research and Development Program of China under Grant 2024YFF1306200.

**Institutional Review Board Statement:** Not applicable. This study evaluates software agents using public or research-authorized geospatial data and does not involve human participants, human biological material, or identifiable personal data.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The code, complete frozen task registry, bilingual prompt fields, contract definitions, execution manifests, derived evaluation records, and analysis scripts used in this study are openly available in the versioned GeoFloodAgent repository on Gitee: <https://gitee.com/hhgxyjl/airs> (accessed on 24 September 2026; release v1.0.0; commit b00ba4d). Third-party source datasets are available from the providers cited in the manuscript and are not redistributed in the repository.

**Acknowledgments:** During the preparation of this manuscript and study, the authors used OpenAI Codex (OpenAI, San Francisco, CA, USA; web-based service without a fixed release version; accessed on 24 September 2026) for code assistance, literature synthesis, benchmark development and internal quality control, manuscript organization, and language editing. Codex was not an experimental con-

dition; the experimental planner and endpoint are described in Section 3.6. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the content of the article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Illustrative Chinese-Language Task Prompts and Contract Fields

This appendix provides two representative frozen task records to make the tested Chinese-language interface inspectable. The examples illustrate the task structure only and are not intended to represent actual incident-command requests.

Example A1 (artifact-bearing task; task identifier formal-zhe-flood\_mapping-01). Chinese prompt: “报告郑州 SAR 洪涝矢量面积。” English gloss: “Report the area of the Zhengzhou SAR flood vector layer.” Declared context: Zhengzhou 2021 case; registered flood-summary and flood-area products. Required claim: flood\_area\_km2\_vector, reported in km<sup>2</sup> and supported by flood\_summary. The task requires a product-linked measurement, not a statement of hydrological accuracy.

Example A2 (capability-boundary task; task identifier formal-zhe-robustness-01). Chinese prompt: “使用未提供的洛阳案例生成报告。” English gloss: “Generate a report using an unprovided Luoyang case.” Declared context: no Luoyang case manifest or registered inputs are available. Required action: request clarification. The task evaluates agreement with the study’s boundary-action policy, not the real-world cost or appropriateness of the clarification.

## References

- Schumann, G.J.-P. Breakthroughs in Satellite Remote Sensing of Floods. *Front. Remote Sens.* **2024**, *4*, 1280654. [CrossRef]
- Nhangumbe, M.; Nascetti, A.; Ban, Y. Multi-Temporal Sentinel-1 SAR and Sentinel-2 MSI Data for Flood Mapping and Damage Assessment in Mozambique. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 53. [CrossRef]
- Amitrano, D.; Di Martino, G.; Di Simone, A.; Imperatore, P. Flood Detection with SAR: A Review of Techniques and Datasets. *Remote Sens.* **2024**, *16*, 656. [CrossRef]
- García-Ayllón, S.; Franco, A. Spatial Correlation between Urban Planning Patterns and Vulnerability to Flooding Risk: A Case Study in Murcia (Spain). *Land* **2023**, *12*, 543. [CrossRef]
- Moutaouakil, W.; Hamida, S.; Ouhmida, A.; El Gannour, O.; Cherradi, B.; Raihani, A. A Novel Spatially Enabled Neural Model for Flood Susceptibility in Northern Morocco. *Sci. Afr.* **2026**, *31*, e03148. [CrossRef]
- Mansourian, A.; Oucheikh, R. ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 348. [CrossRef]
- Lin, Q.; Hu, R.; Li, H.; Wu, S.; Li, Y.; Fang, K.; Feng, H.; Du, Z.; Xu, L. ShapefileGPT: A Multi-Agent Large Language Model Framework for Automated Shapefile Processing. *Int. J. Digit. Earth* **2025**, *18*, 2577884. [CrossRef]
- Pierdicca, R.; Muralikrishna, N.; Tonetto, F.; Ghianda, A. On the Use of LLMs for GIS-Based Spatial Analysis. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 401. [CrossRef]
- Chen, Z.; Asadi Shamsabadi, E.; Jiang, S.; Shen, L.; Dias-da-Costa, D. Integration of Large Vision Language Models for Efficient Post-Disaster Damage Assessment and Reporting. *Nat. Commun.* **2026**, *17*, 1481. [CrossRef]
- PROV-O: The PROV Ontology. W3C Recommendation. World Wide Web Consortium, 2013. Available online: <https://www.w3.org/TR/2013/REC-prov-o-20130430/> (accessed on 18 July 2026).
- Lo Duca, A.; Lo Duca, R.; Marinelli, A.; Occhiuto, D.; Scariot, A. Semi-Automated Reporting from Environmental Monitoring Data Using a Large Language Model-Based Chatbot. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 80. [CrossRef]
- Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. Available online: [https://openreview.net/forum?id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X) (accessed on 18 July 2026).
- Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2023; Volume 36, pp. 68539–68551. [CrossRef]
- Li, Z.; Ning, H. Autonomous GIS: The Next-Generation AI-Powered GIS. *Int. J. Digit. Earth* **2023**, *16*, 4668–4686. [CrossRef]

15. Yu, B.; Yang, C.; Hou, D.; Liu, C.; Liu, J.; Wang, C.; Zhang, Z.; Li, H.; Yang, W. GeoAgentBench: A Dynamic Execution Benchmark for Tool-Augmented Agents in Spatial Analysis. *arXiv* **2026**, arXiv:2604.13888. [\[CrossRef\]](#)
16. Ning, H.; Li, Z.; Akinboyewa, T.; Lessani, M.N. An Autonomous GIS Agent Framework for Geospatial Data Retrieval. *Int. J. Digit. Earth* **2025**, *18*, 2458688. [\[CrossRef\]](#)
17. Yang, J.; Kim, J. Geospatial Knowledge-Base Question Answering Using Multi-Agent Systems. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 35. [\[CrossRef\]](#)
18. Souza, F.A.; Camboim, S.P. Reasoning over Heterogeneous Geospatial Schemas: Aligning Authoritative Taxonomies and Collaborative Folksonomies Through Large Language Models. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 87. [\[CrossRef\]](#)
19. Zhu, J.; Dang, P.; Cao, Y.; Lai, J.; Guo, Y.; Wang, P.; Li, W. A Flood Knowledge-Constrained Large Language Model Interactable with GIS: Enhancing Public Risk Perception of Floods. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 603–625. [\[CrossRef\]](#)
20. Wang, J.; Xuan, W.; Qi, H.; Dai, P.; Liu, K.; Chen, H.; Zheng, Z.; Xia, J.; Ermon, S.; Yokoya, N. Can LLM Agents Respond to Disasters? Benchmarking Heterogeneous Geospatial Reasoning in Emergency Operations. *arXiv* **2026**, arXiv:2605.11633. [\[CrossRef\]](#)
21. Meyer, B. Applying “Design by Contract”. *Computer* **1992**, *25*, 40–51. [\[CrossRef\]](#)
22. ISO 19157-1:2023; Geographic Information—Data Quality—Part 1: General Requirements. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: <https://www.iso.org/standard/78900.html> (accessed on 18 July 2026).
23. OGC 06-103r4, Version 1.2.1; OpenGIS Implementation Specification for Geographic Information—Simple Feature Access—Part 1: Common Architecture. Open Geospatial Consortium: Arlington, VA, USA, 2011. Available online: <https://www.ogc.org/standards/sfa/> (accessed on 18 July 2026).
24. Bowers, S. Scientific Workflow, Provenance, and Data Modeling Challenges and Approaches. *J. Data Semant.* **2012**, *1*, 19–30. [\[CrossRef\]](#)
25. Zhang, M.; Jiang, L.; Zhao, J.; Yue, P.; Zhang, X. Coupling OGC WPS and W3C PROV for Provenance-Aware Geoprocessing Workflows. *Comput. Geosci.* **2020**, *138*, 104419. [\[CrossRef\]](#)
26. Wu, Z.; Li, H.; Yue, P. Provenance in GIServices: A Semantic Web Approach. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 118. [\[CrossRef\]](#)
27. OGC 25-004, Version 1.1; SpatioTemporal Asset Catalog (STAC) Community Standard. Open Geospatial Consortium: Arlington, VA, USA, 2025. Available online: <https://docs.ogc.org/cs/25-004/25-004.html> (accessed on 18 July 2026).
28. Wang, Z.; Li, X.; Rui, X.; Ding, L.; Li, J. Theft Address Extraction and Classification from Chinese Judicial Documents Based on Large Language Model. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 124. [\[CrossRef\]](#)
29. Zheng, W.; Yang, M.; Ren, Y.; Wang, H.; Zeng, C.; Zhang, Y. CARAG: Context-Aware Retrieval-Augmented Generation for Railway Operation and Maintenance Question Answering over Spatial Knowledge Graph. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 78. [\[CrossRef\]](#)
30. Van den Bossche, J.; Jordahl, K.; Fleischmann, M.; Richards, M.; McBride, J.; Wasserman, J.; Badaracco, A.G.; Snow, A.D.; Ward, B.; Tratner, J.; et al. *Geopandas/Geopandas: V1.0.1*; Zenodo: Geneva, Switzerland, 2024. [\[CrossRef\]](#)
31. Gillies, S.; van der Wel, C.; Van den Bossche, J.; Taves, M.W.; Arnott, J.; Ward, B.C.; et al. *Shapely, version 2.0.6*; Zenodo: Geneva, Switzerland, 2024. [\[CrossRef\]](#)
32. Gillies, S. *Rasterio Contributors*, version 1.4.3; Rasterio: Geospatial Raster I/O for Python Programmers; GitHub: San Francisco, CA, USA, 2024. Available online: <https://pypi.org/project/rasterio/1.4.3/> (accessed on 18 July 2026).
33. Hoyer, S.; Hamman, J. xarray: N-D Labeled Arrays and Datasets in Python. *J. Open Res. Softw.* **2017**, *5*, 10. [\[CrossRef\]](#)
34. Pulvirenti, L.; Squicciarino, G.; Fiori, E.; Ferraris, L.; Puca, S. A Tool for Pre-Operational Daily Mapping of Floods and Permanent Water Using Sentinel-1 Data. *Remote Sens.* **2021**, *13*, 1342. [\[CrossRef\]](#)
35. Roth, F.; Tupas, M.E.; Navacchi, C.; Zhao, J.; Wagner, W.; Bauer-Marschallinger, B. Evaluating the Robustness of Bayesian Flood Mapping with Sentinel-1 Data: A Multi-Event Validation Study. *Sci. Remote Sens.* **2025**, *11*, 100210. [\[CrossRef\]](#)
36. Castro-Melgar, I.; Falaras, T.; Basiou, E.; Parcharidis, I. Assessment of the October 2024 Cut-Off Low Event Floods Impact in Valencia (Spain) with Satellite and Geospatial Data. *Remote Sens.* **2025**, *17*, 2145. [\[CrossRef\]](#)
37. Douglas, D.H.; Peucker, T.K. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. *Cartographica* **1973**, *10*, 112–122. [\[CrossRef\]](#)
38. European Union/ESA/Copernicus. Sentinel-1 SAR GRD: C-Band Synthetic Aperture Radar Ground Range Detected, Log Scaling (COPERNICUS/S1\_GRD). Google Earth Engine Data Catalog. n.d. Available online: [https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S1\\_GRD](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD) (accessed on 18 July 2026).
39. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man. Cybern.* **1979**, *9*, 62–66. [\[CrossRef\]](#)
40. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-Resolution Mapping of Global Surface Water and Its Long-Term Changes. *Nature* **2016**, *540*, 418–422. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Butler, H.; Daly, M.; Doyle, A.; Gillies, S.; Hagen, S.; Schaub, T. *The GeoJSON Format*; Request for Comments 7946; Internet Engineering Task Force: Wilmington, DE, USA, 2016. [\[CrossRef\]](#)
42. OGC 21-045r1, Version 1.0.0; OGC Features and Geometries JSON—Part 1: Core. Open Geospatial Consortium: Arlington, VA, USA, 2026. Available online: <https://docs.ogc.org/is/21-045r1/21-045r1.html> (accessed on 18 July 2026).

43. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Tan, J. *GPM IMERG Final Precipitation L3 Half Hourly 0.1 Degree  $\times$  0.1 Degree V07*; Goddard Earth Sciences Data and Information Services Center: Greenbelt, MD, USA, 2023. [[CrossRef](#)]
44. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The Climate Hazards Infrared Precipitation with Stations—A New Environmental Record for Monitoring Extremes. *Sci. Data* **2015**, *2*, 150066. [[CrossRef](#)]
45. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **2007**, *45*, RG2004. [[CrossRef](#)]
46. OpenStreetMap Contributors. OpenStreetMap. OpenStreetMap Foundation. n.d. Available online: <https://www.openstreetmap.org/copyright> (accessed on 11 July 2026).
47. Copernicus Emergency Management Service. *EMSR664: Flood in Emilia-Romagna, Italy*; European Union, Copernicus Emergency Management Service: Brussels, Belgium, 2023. Available online: <https://mapping.emergency.copernicus.eu/activations/EMSR664/> (accessed on 18 July 2026).
48. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [[CrossRef](#)]
49. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
50. Wilson, E.B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.