TECHNICAL REPORT / SYSTEM 02

Atlas Research Lab

Making multi-agent research visible, inspectable, and honest about its limits.

Fatima Aguilar Decision Systems Lab July 2026

Live app: atlas-research-lab.fsaguilar16.chatgpt.site
Source: github.com/fa366193/atlas-research-lab

Abstract

A visible laboratory for agentic research

Atlas Research Lab is an interactive prototype for observing how a verification-focused team of autonomous agents might investigate a question. Five specialized characters make the research process legible: a planner decomposes the question, a researcher gathers evidence, a skeptic attacks assumptions, an analyst tests the result, and a replicator independently reconstructs the reasoning chain.

The current public release is deliberately framed as a simulation. It adapts the interface, evidence themes, trace, and proposed next experiment to a visitor-authored question, but it does not claim to have retrieved or verified live sources. This distinction is central to the project: interface polish must not be mistaken for epistemic reliability.

The contribution is therefore twofold: an interaction model that makes agent roles physically understandable, and a research roadmap for turning that model into an evaluated autonomous investigation system.

Research question

Can verification-focused agent roles make autonomous research easier to inspect and, once connected to real tools and evaluation, more reliable?

Current deliverable

Surface	Function	Status
Question composer	Visitors initiate a run	Implemented
Animated studio	Roles and stages become observable	Implemented
Audit trace	Shows a complete simulated action log	Implemented
Live source retrieval	Collects and verifies real evidence	Research phase
Atlas-100	Held-out reliability benchmark	Proposed

System architecture

Five roles, one inspectable chain

Atlas treats research as a sequence of distinct responsibilities rather than a single opaque completion. Separating roles makes failure easier to locate and creates natural intervention points for evaluation.

1. Planner

Pip converts a broad visitor question into candidate explanations, measurable outcomes, and a source plan. The planner also decides when evidence is insufficient and a new investigation loop is required.

2. Researcher

Scout gathers primary sources and datasets, records provenance, and identifies missing coverage. In a production system, this role would use constrained search tools and store evidence as immutable records.

3. Skeptic

Mochi is optimized for productive disagreement. The role seeks alternate explanations, measurement errors, selection effects, and claims whose confidence exceeds their support.

4. Analyst

Byte turns evidence into reproducible calculations. Analyses should execute in a sandbox, record dependencies and random seeds, and expose sensitivity to alternate assumptions.

5. Replicator

Dot receives the question, source set, and method without the original result. Agreement is useful only when the independent run can reconstruct the chain and explain any divergence.

The visible studio is not decoration added after the architecture. It is the architecture's public explanation layer: each spatial movement corresponds to a change in research responsibility.

Interaction design

From a fixed demo to a visitor-authored investigation

The final prototype starts with a question composer. Submitting a question resets the run, classifies broad evidence themes, updates the studio board, and begins a six-stage animation. A visitor may pause, replay, restart, jump to any stage, or inspect the full trace.

Six-stage playback

01	**Briefing**	Question decomposition and hypothesis formation
02	**Gathering**	Source planning and evidence collection
03	**Challenging**	Counter-evidence and assumption testing
04	**Analyzing**	Robustness and sensitivity checks
05	**Replicating**	Independent reconstruction
06	**Outcome**	Bounded recommendation with limitations

Honest simulation boundary

The outcome card says 'simulated outcome - requires real-world verification.' The trace describes intended research operations rather than fabricated citations. This protects the core design claim while avoiding the false impression that an animated interface has completed empirical research.

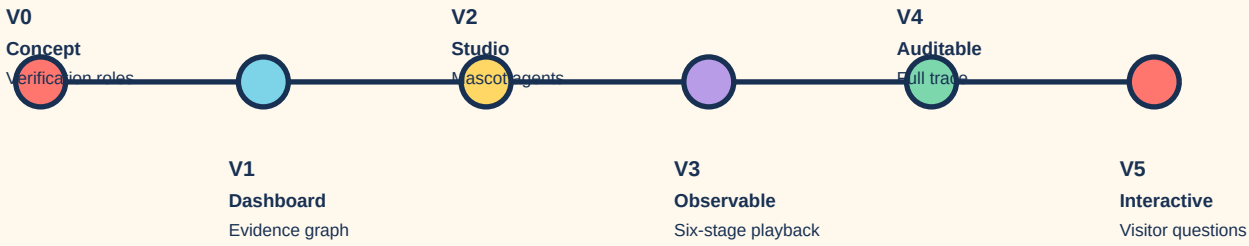
Accessibility and control

The experience uses semantic form controls, visible labels, keyboard-operable buttons, modal dialog semantics, reduced information density on small screens, and explicit pause/restart controls. Future work should add a reduced-motion mode and screen-reader announcements for stage transitions.

Development notebook

The progression of the project

Atlas did not emerge fully formed. Each revision responded to a concrete weakness observed in the version before it.



V0 - Concept. The initial idea was an open multi-agent research laboratory with planner, researcher, skeptic, analyst, citation auditor, and replication roles. The core insight was that admissions-grade work requires an evaluated research claim, not merely an agent demo.

V1 - Evidence dashboard. The first implementation used a dark green scientific interface, a live trace, and an evidence graph. It communicated seriousness but resembled an existing project and the playback changed too little to feel alive.

V2 - Research studio. The visual system moved to warm cream, coral, sky blue, yellow, lavender, and mint. Each role became a mascot inside a dimensional room, turning abstract orchestration into physical action.

V3 - Observable workflow. Playback expanded into six timed stages. Mascots moved, tools animated, speech bubbles explained work, and a final result appeared only after replication.

V4 - Auditable interaction. The full-trace button became a working modal. Action bubbles moved clearly above the mascots. Visitors could pause, restart, or jump directly to a stage.

V5 - Visitor-authored questions. A question composer replaced the single preselected prompt. Broad evidence themes, the trace, and the bounded outcome now adapt to the visitor's question while remaining visibly labeled as simulation.

Implementation

Technical decisions

Frontend

The application uses React state for the active question, draft question, playback stage, playback state, and trace visibility. A timed effect advances stages every 2.3 seconds and stops after the outcome. Deterministic classification maps common domains - climate, education, transportation, and health - to relevant evidence themes and bounded recommendation templates.

Visual system

The studio, mascots, furniture, evidence board, floor perspective, speech bubbles, and motion are rendered in HTML and CSS. This keeps the primary experience interactive, responsive, accessible, and lightweight. The separate social-preview card uses a 3D clay-like illustration consistent with the interface palette.

Deployment

The application is built with a Next.js-compatible app structure and vinext, producing a Cloudflare Workers-compatible deployment. Source is versioned publicly on GitHub under the MIT License.

State model

State	Purpose
draftQuestion	Editable composer input
question	Committed question controlling the run
stage	Integer from 0 to 5 controlling scene and timeline
playing	Starts or pauses timed progression
traceOpen	Controls the audit dialog

The prototype intentionally avoids a backend because it does not yet perform real investigation. Adding persistence or an API before the research execution model is defined would create misleading product depth without scientific depth.

Evaluation roadmap

From interaction prototype to research contribution

The next milestone is Atlas-100: a held-out benchmark of one hundred urban-climate and civic research questions with expert-reviewed evidence requirements, reproducible reference analyses, and explicit uncertainty.

Primary metrics

Citation precision

Fraction of citations that directly support the associated claim.

Claim support

Fraction of material claims justified by evidence or explicit inference.

Replication success

Fraction of results independently reconstructed within tolerance.

Calibration

Agreement between stated confidence and observed correctness.

Coverage

Fraction of important counter-evidence and limitations identified.

Cost and latency

Resources required per reliable investigation.

Ablation studies

Run the same questions with the skeptic removed, the replicator removed, shared memory disabled, role prompts collapsed into one agent, and fixed versus adaptive planning. The central empirical claim should be supported only if verification roles produce statistically meaningful improvements on held-out questions.

Training plan

Store structured trajectories containing question, plan, actions, evidence, claims, confidence, critiques, revisions, and outcome grades. Begin with supervised learning from successful traces, then train routing and stopping policies from preference data. Preserve a locked test set and report negative results.

Limitations and ethics

What Atlas cannot yet claim

The public prototype does not browse the web, inspect primary sources, run code, or validate its own outcome. Its domain classification is heuristic. The displayed metrics are research targets and interface content, not results from a completed Atlas-100 study.

Key risks

Risk	Mitigation
Persuasive animation creates false trust	Persistent simulation labels and source-level auditability
Agents amplify the same model bias	Independent models, role isolation, and diverse human review
Citation laundering	Claim-level entailment checks and immutable source snapshots
Overconfident synthesis	Calibration scoring and explicit abstention policies
Sensitive civic recommendations	Community review and distributional-impact analysis
Benchmark overfitting	Locked cases, rotating test sets, and external replication

Design principle

The system should become less confident as uncertainty grows, not more verbose. A delightful interface is valuable only when it makes epistemic status clearer.

Conclusion

A lab notebook, not a finished claim

Atlas demonstrates a new way to explain agentic research: make specialized responsibilities visible, let visitors control the investigation, and attach a legible trace to the outcome. The evolution from a fixed evidence dashboard to an interactive mascot studio improved both originality and comprehension.

The most important decision was to preserve the boundary between simulation and evidence. That boundary gives the project a credible path forward. The next release should not simply add more animation; it should connect each visible action to a real tool call, a durable source record, an executable analysis, and a measurable evaluation.

The long-term research contribution will be established when Atlas can answer a narrower question with evidence: whether explicit skepticism and independent replication improve autonomous research reliability enough to justify their additional cost.

Project links

Live application
https://atlas-research-lab.fsaguilar16.chatgpt.site

Public source repository
https://github.com/fa366193/atlas-research-lab

License
MIT License