

ARIS: Autonomous Research via Adversarial Multi-Agent Collaboration

Ruofeng Yang

Shanghai Jiao Tong University
Supervisor: Shuai Li

VALUE 2026



Research Paradigm with Stronger Agent

As the long-horizon agent becomes stronger and stronger

- It seems that each components of research can be done with current agent with long-horizon task

The answer is **Definitely None** for Auto-research with high precision requirements.

We can generate a **A+B** paper that **looks correct** within an hour

扫文献

找Idea

跑实验 / baseline

写 paper / rebuttal

Resubmit

PPT/Talk Draft

The Failure Mode That Hides

Dominant failure mode of long-horizon agent research is **not** visible breakdown.
It is **plausible unsupported success**:

- Results may be real yet **misreported**
- Claims may **outrun the evidence** that licenses them
- Readers may **silently inherit** the executor's framing

Visible failure → easy to fix.

Plausible unsupported success → silently propagates.

Three Typical Manifestations

Model-Derived Reference Labels

Agent uses its own outputs as ground truth for self-evaluation.

WHAT TO LOOK FOR

eval pipeline reads model output as label source

Self-Normalized Scores

Metric denominator comes from the model's own predictions, inflating performance.

WHAT TO LOOK FOR

metric denominator uses model probability

Phantom Results

Claimed numbers in the manuscript do not match raw output files.

WHAT TO LOOK FOR

trace every number back to a JSON / CSV file

These are not corner cases — they are high-frequency.

The Stringent Assumption

*Any long-term task performed by a single agent is **unreliable**.*

*We must divide the workflow into sub-workflows, and use a **cross-family** reviewer to audit the output at each step **independently**.*

Why so strict? Two analogies

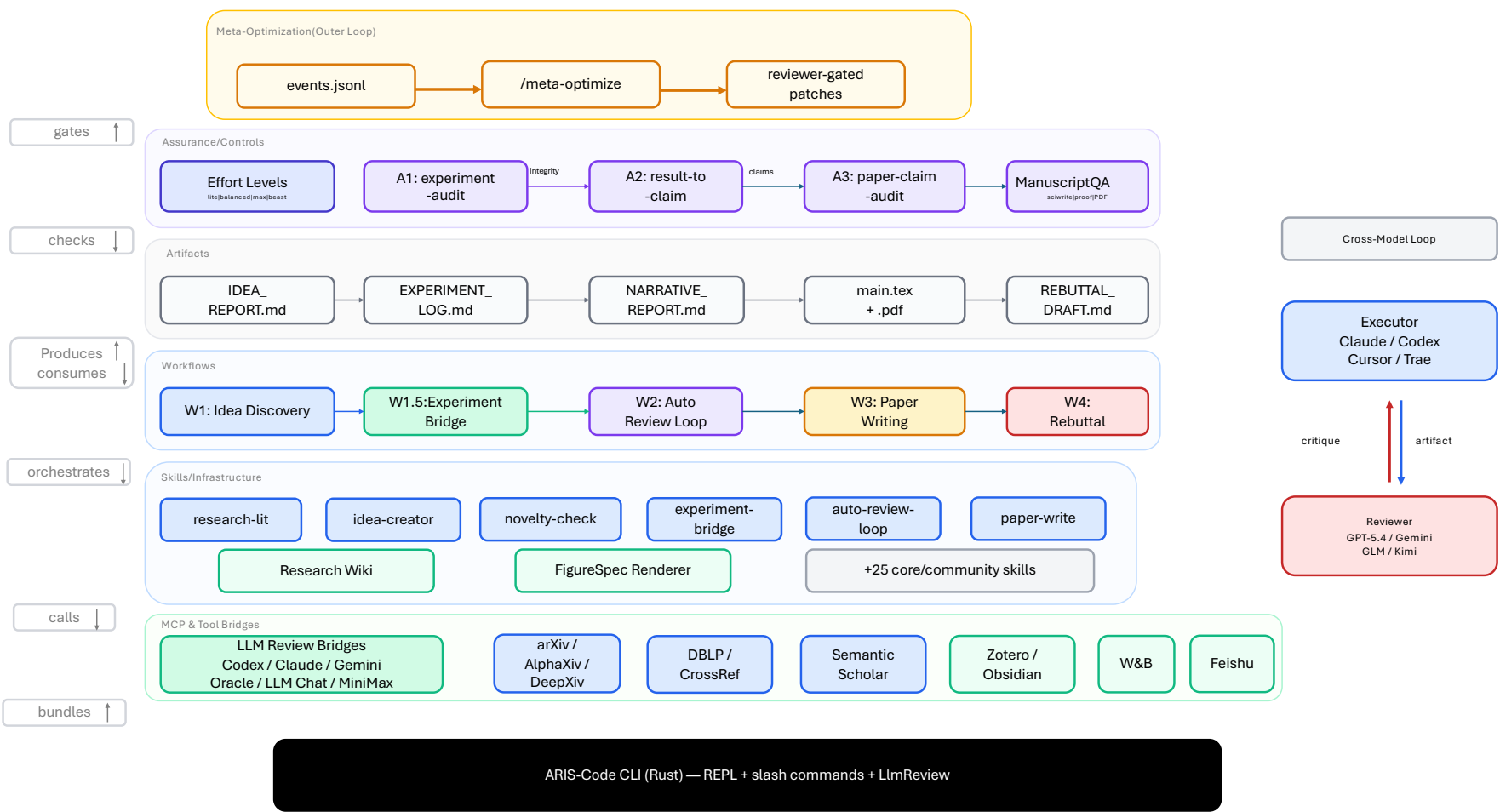
Adversarial vs Stochastic Bandits

- self-review = stochastic (predictable noise)
- cross-family = adversarial: probes weak spots
- *adversarial feedback is harder to game*

2-player Nash

- executor + reviewer breaks self-play blind spots
- n-player coordination has diminishing returns
- default $1 + 1 = 2$: just enough

Overview of ARIS: 3-Layer Architecture

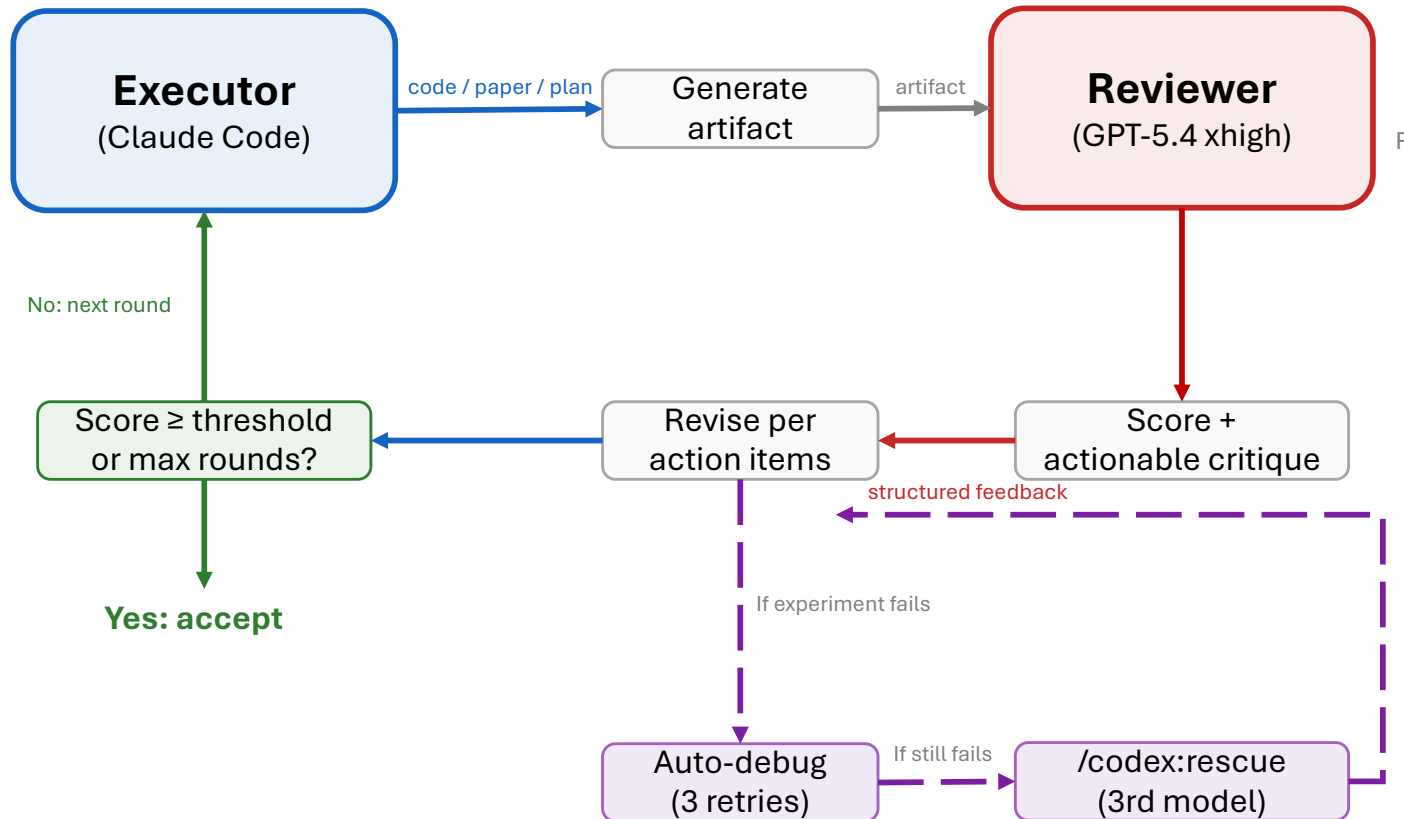


Execution
65+ skills · MCP bridges
figures · persistent wiki

Orchestration
5 workflows

Assurance
3-stage cascade · MS checks
meta-opt outer loop

Cross-Model Adversarial Loop



Information Assurance

- **Access:** doc-only /log-only
- If no absolute path is provided, the reviewer will **directly reject this chat**.

Fresh Memory for Reviewer

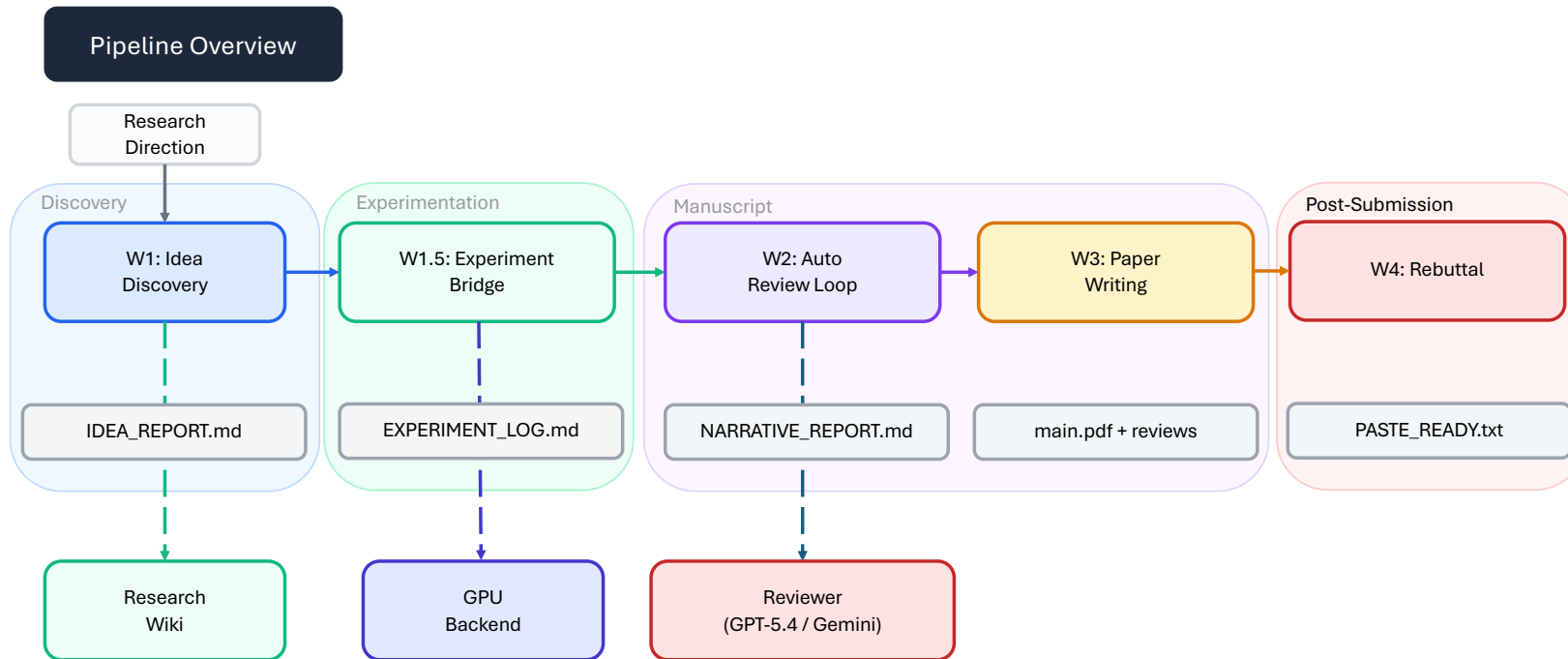
- To avoid memory decay for multi round review

Accumulate Memory for Debugger

- Debug need context

Default for assurance: cross-family + fresh thread.

Overview of Workflow



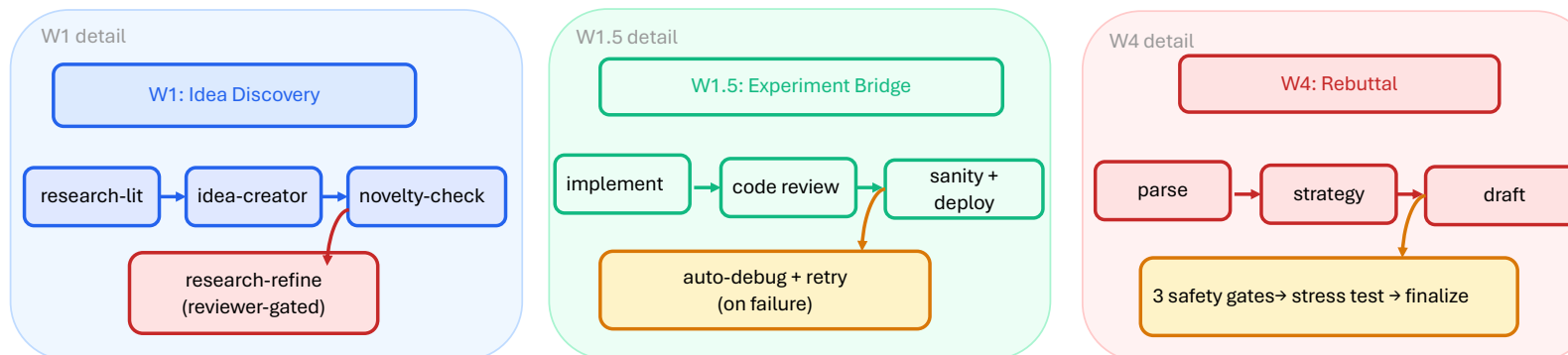
Highly flexible

- Each workflow independently instead of end2end
- **human can in the loop**

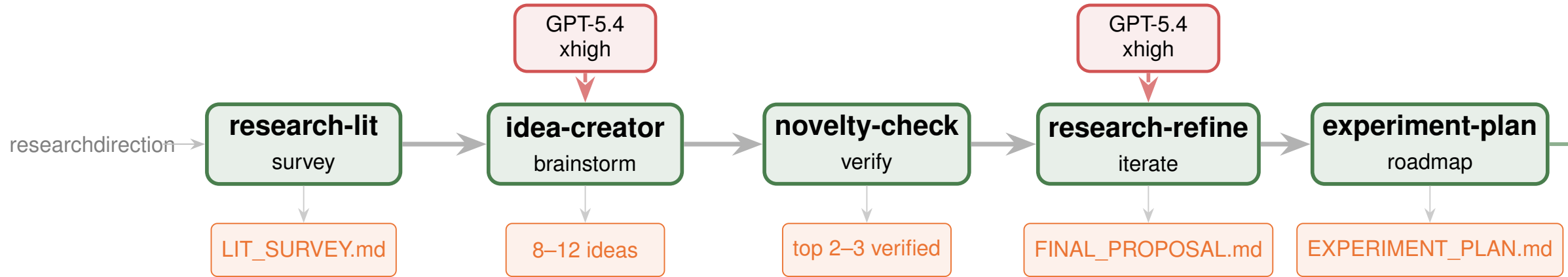
Cover Full Pipeline

- More workflow:
 - Resubmit workflow
 - Talk workflow

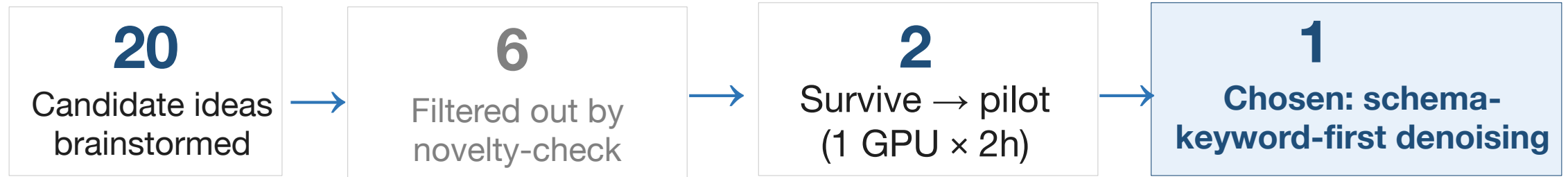
Workflow Internals (W2→ Fig.~6, W3 → Fig.~7)



W1: From Brainstorm to Research Question



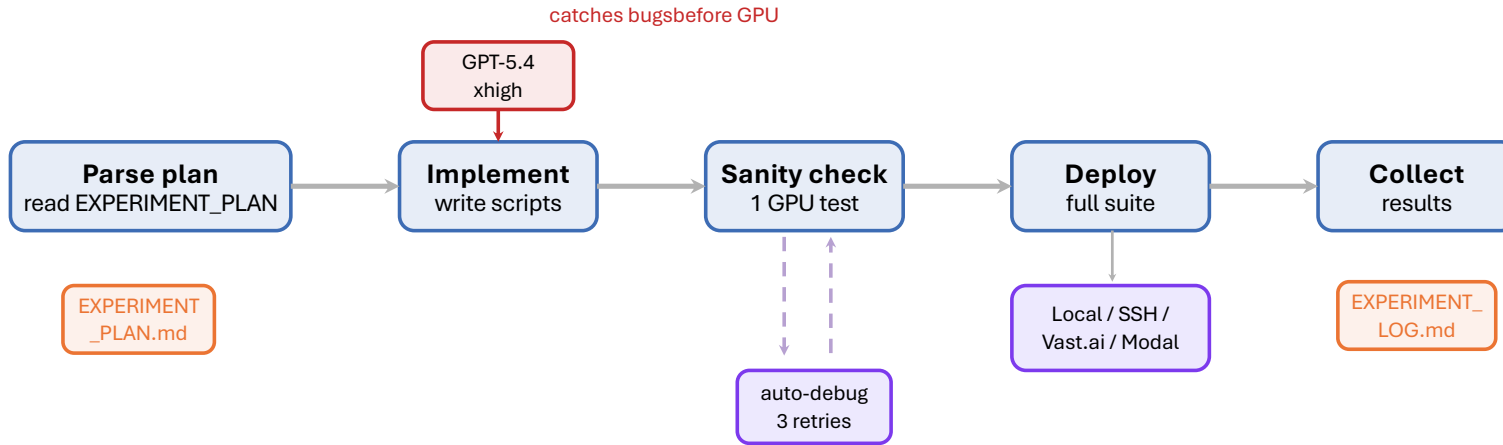
RUNNING EXAMPLE M0 – DLLM × STRUCTURED GENERATION



HONEST LIMITATION

idea-creator is *intentionally conservative* — it often proposes plausible “A + B” combinations and does not yet replace expert research taste.

W1.5: Experiment Bridge



MUST LOG and Monitor

- To guarantee experiments really conducted

M1 RESEARCH QUESTION

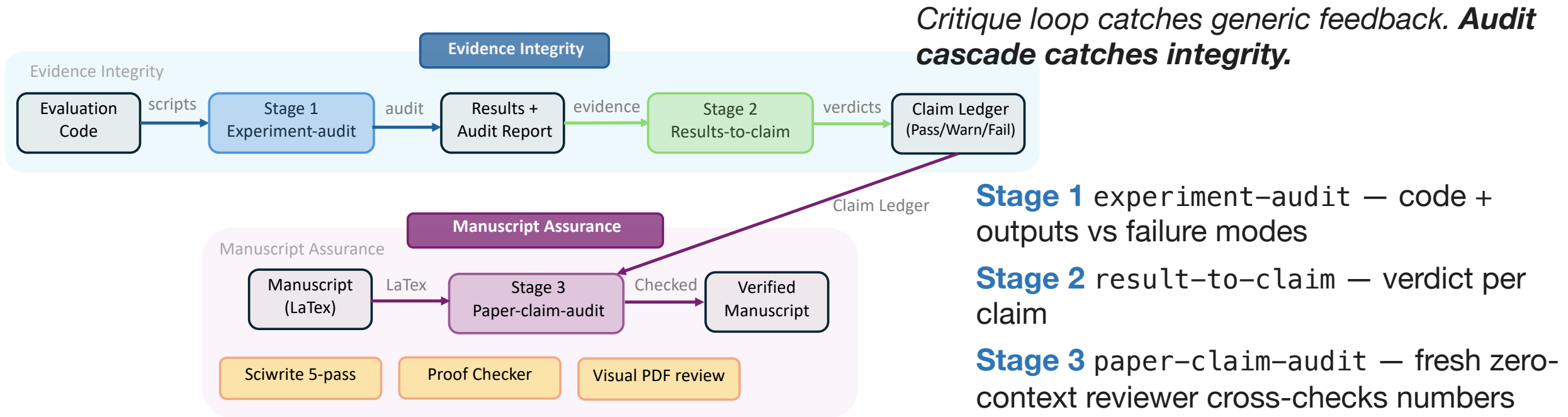
Does **schema-keyword-first denoising** beat random masking on exact-parse success rate under *external* jsonschema validation, for DLLM under tree-shaped JSON schema (depth 2-4)?

M2 EXPERIMENT_PLAN.md (auto-gen):

- Dataset: 164 schema-constrained records
- Model: ~200M DLLM
- Baselines: random-mask, left-to-right AR
- Primary: **raw json.loads + jsonschema**
- Compute: 4xA100 · 24h × 3 seeds · 27 jobs

注意“独立”关键 — plan 是对的, bug 一会儿会出在 execution wrapper 偷偷偏离这个 spec。

Audit Cascade (3-stage Audit)



► Running example M3:

Initial result: schema-aware **+6.2** success rate over random-mask baseline.

Stage 1 audit (Codex GPT-5.5 xhigh, fresh thread) flagged: *evaluator silently used pipeline's JSON sanitizer that default-fills malformed keys.*

Re-evaluated with raw `json.loads + jsonschema`: **gap collapses to +1.4.**

Verdict: `WARN_corrected` — effect real but scope narrowed. **Bug caught before claim amplification.**

Manuscript Assurance + Beyond Submission

4 Draft Checks

Pass A

5-pass sci editing

*clutter · voice · structure
· terms · numerical*

Pass B

proof-checker

*20-cat taxonomy + red
team*

Pass C

kill-argument

*fresh adversarial reviewer
with negative prompt*

Pass D

citation-audit

*existence · metadata ·
context*

M4 ledger entry

```
claim_id=C7
evidence=results/json_2026_05_06.json
integrity=WARN_corrected
scope="under external jsonschema only"
```

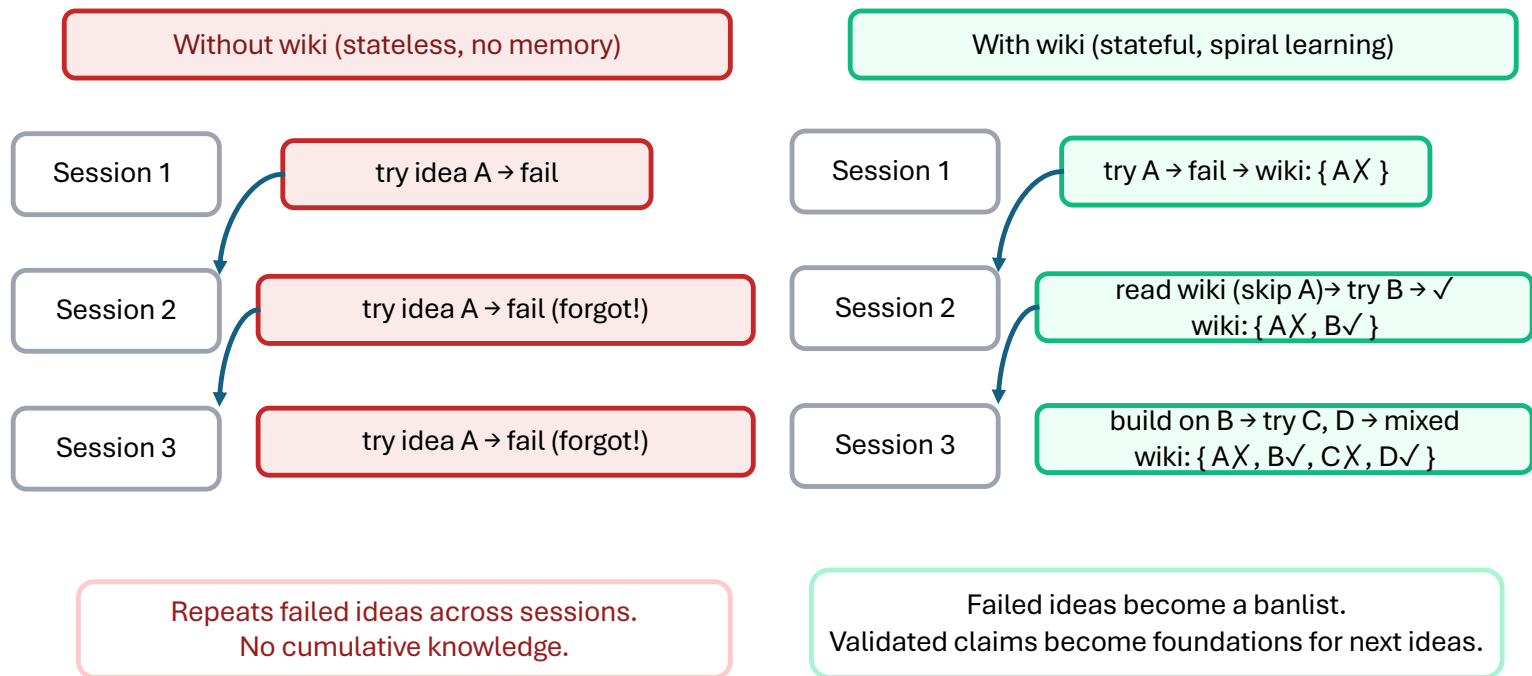
M5 MANUSCRIPT PROSE

“Under external schema validation (Table 3), schema-aware denoising improves exact-parse success by **+1.4** over random masking (95% CI ± 0.8); an additional **+4.8** measured under in-pipeline sanitization is an evaluator self-repair artifact (Appendix B.2).”

ARIS verdict aesthetic: **诚实但不杀稿** — claim narrows from “X is best” to “X is best when external validator used”.

Beyond submission: W4 rebuttal + W5 resubmit · feeds wiki for next cycle

How about Continual Learning: Research Wiki



*Idea Discovery 先广度优先搜索,
Following workflow 再深度优先搜索,
Research wiki 再回到广度优先搜索*

Research wiki Design:

- Node: (Paper, idea, experiment, claim)
- Paper link with graph

Round 1: read 15 papers → wiki remembers → idea A → experiment → FAIL
wiki records: "A fails because OOM at batch>32, loss diverges"

Round 2: /idea-creator reads wiki → sees A failed → generates idea D (avoids A's trap)
→ experiment → PARTIAL SUCCESS
wiki records: "D works on small models, fails on large"

Round 3: /idea-creator reads wiki → knows A failed + D partial → generates idea F
(combines D's success with new approach) → experiment → SUCCESS 🎉

Automation \neq Outsourcing Research Taste

ARIS automates *labor*, not *judgment*.

ARIS handles

- Literature scan
- Some “A+B” idea via idea discovery and research wiki, even **maybe SoTA**
- Experiment ops + GPU deployment
- Paper drafting + figure rendering
- Rebuttal logistics

Human owns

- **Problem framing** — what question matters
- **Evidence threshold** — when is enough
- **Claim fairness** — warranted or stretched
- **Research taste** — built by judgment

For PhD students: don't outsource the parts that train you.

A Reliable Hallucination is Effective for Idea

In this share, we often emphasize the importance of **honesty**. However, *Hallucination is also important* for idea discovery

**Creation Agent
(Hallucination)**



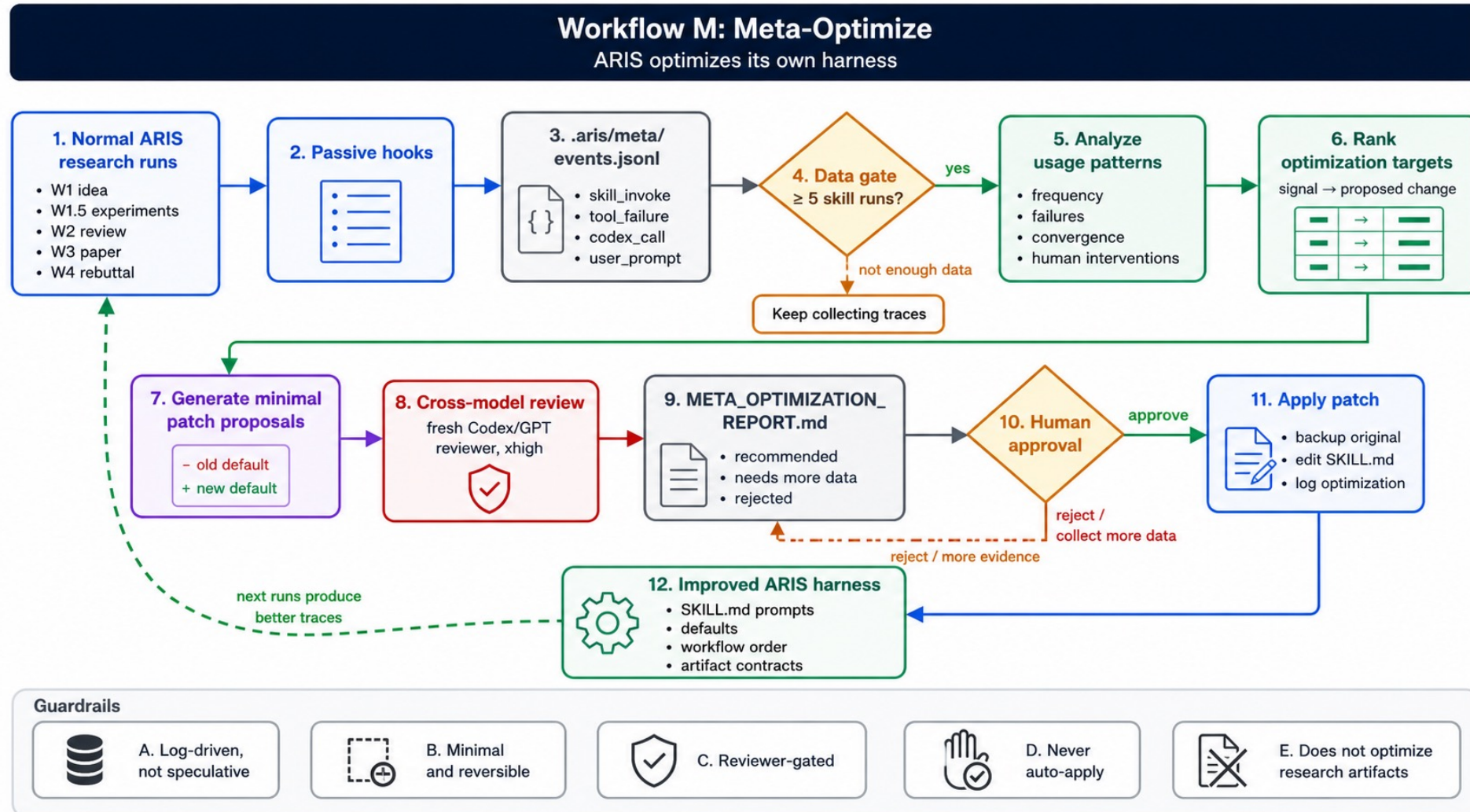
Adversarial Supervisor

Honesty Agent

- For example, Gemini 3.1-Pro is helpful in the Brainstorming with GPT 5.5 Pro and Claude 4.7 Opus
- We also support any other agent models combination

How about Continual Learning: Meta-Optimization

Design Yourself ARIS During running workflow with different models



Future work

Self-Improving Foundation Models

Apply auto-research to architectural innovation:

- dense \rightarrow MoE
- gated attention discovery
- attention-sink mitigation

Goal: ARIS detects inefficiency \rightarrow proposes architectural change \rightarrow pilots \rightarrow validates before full training.

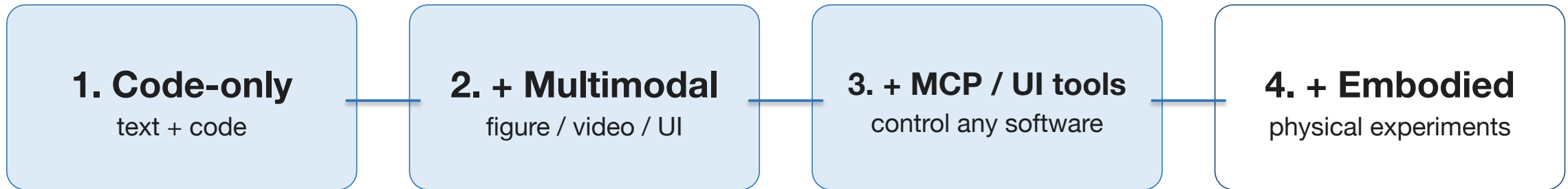
Auto-Research is a clear task with variable reward (diverse Benchmark) for Long-horizon Agent

- Similar position with **Math Task for previous agent**

Goal: Post-training Long-horizon agent with auto-research task

Future work: Go beyond Coding

ARIS today: text/code/artifact workflows only. **No embodied experiments.**



*World model + action grounding → **embodied research agents** is plausible.*

跨学科的终点很可能是 embodied research agent — 今天还远, trajectory 清晰。



HuggingFace
Welcome to upvote& Star

Meta-demo:

This talk deck and tech report were produced with **ARIS in the loop.**

Human taste, agent labor, adversarial audit.

Thanks!



Wechat



Ruofeng Yang

4-th PhD candidate

- Shanghai Jiao Tong University
- Research: Video/image Generation, Agent, Auto Research
- <https://wanshuiyin.github.io/>
- **Open to the Job Market**



Yongcan Li

1-th PhD

- Shanghai Jiao Tong University
- Research: Analysis of Generative Models, Agent, Auto Research



Supervisor: Shuai Li

- Associate Professor
- Shanghai Jiao Tong University
- Research: RL/ML theory
- <https://shuaili8.github.io/>