NeurIPS 2026

Lemon

Evidence-Risk-Aware Adaptive Organization for Long-Horizon LLM Agents



Zimo Yin[1], Haipeng Jiang[1], Kailong Ren[1], Zhetao Sun[1], Guangyi Lv[1], Congli Yin[1]
Likang Wu[2], Hongke Zhao[2], Ming He[1], Peng Wang[1], Jianping Fan[3]

[1] AI Lab, Lenovo Research [2] Tianjin University [3] Fudan University

Agents fail when they lose track of evidence

01

Missing support

Too little evidence to support a reliable answer.

02

Unresolved contradictions

Conflicting sources remain unreconciled.

03

Evidence lost under context pressure

Repeated summaries drop the decisive detail.

Typical agent workflow

Plan → Tools → Memory

Evidence risk stays implicit

Evidence risk guides execution

State signals

Evidence gaps and contradictions
Context pressure
Retrieved memory
Tools and remaining budget

→

Evidence-risk state

Conditions the next action

→

Execution decisions

Route tasks
Call tools
Expand workers
Verify claims
Compress context
Use memory

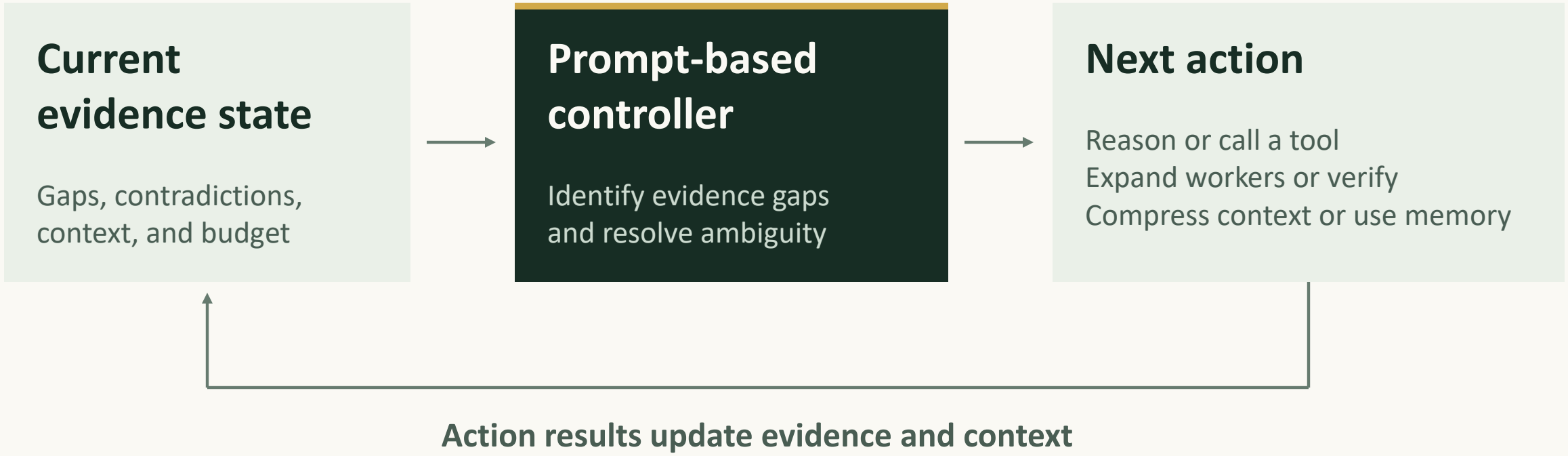
Within one agent

The current state guides each execution step.

Across agents

Agents exchange proposals with evidence, critique claims, and reach claim-level consensus.

The single-agent control loop

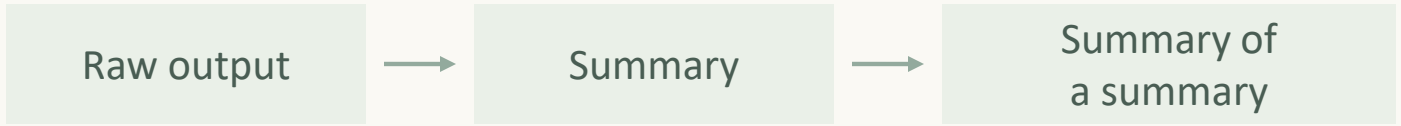


Code enforces character thresholds and round limits.

Recoverable and reusable evidence

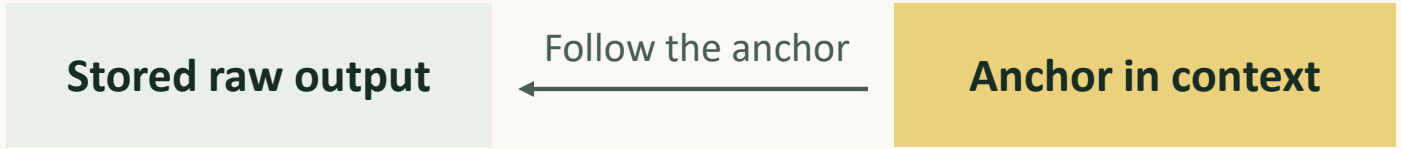
Recoverable context

Forward-only summarization



Decisive detail can be lost

Lemon stores the raw output and leaves an anchor



Later steps can revisit the original evidence.

Self-evolving memory

Retains reusable process fragments, such as tool heuristics.

Specialized MCP tools

Reach fine-grained visual and street-view evidence.

GAIA accuracy and token efficiency

GAIA test leaderboard

91.36%

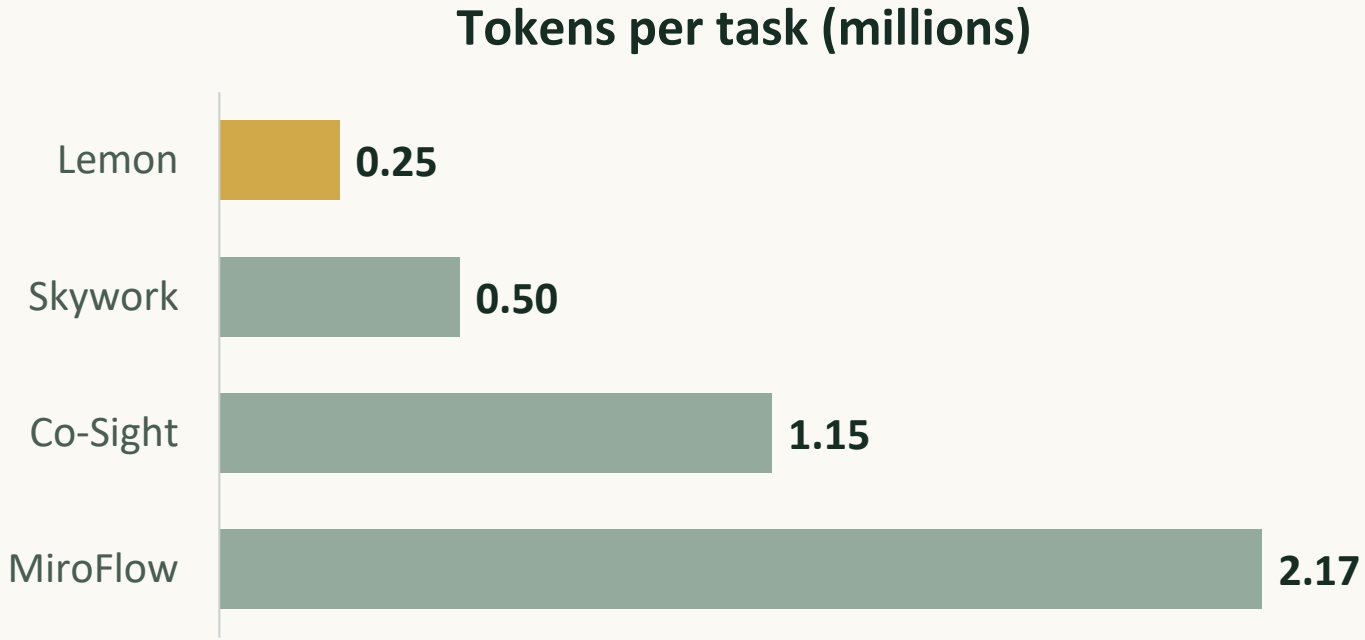
275 of 301 tasks

Submission dated 6 Feb 2026

xbench-DeepSearch-2510

80 / 100

In-house run under a separate protocol



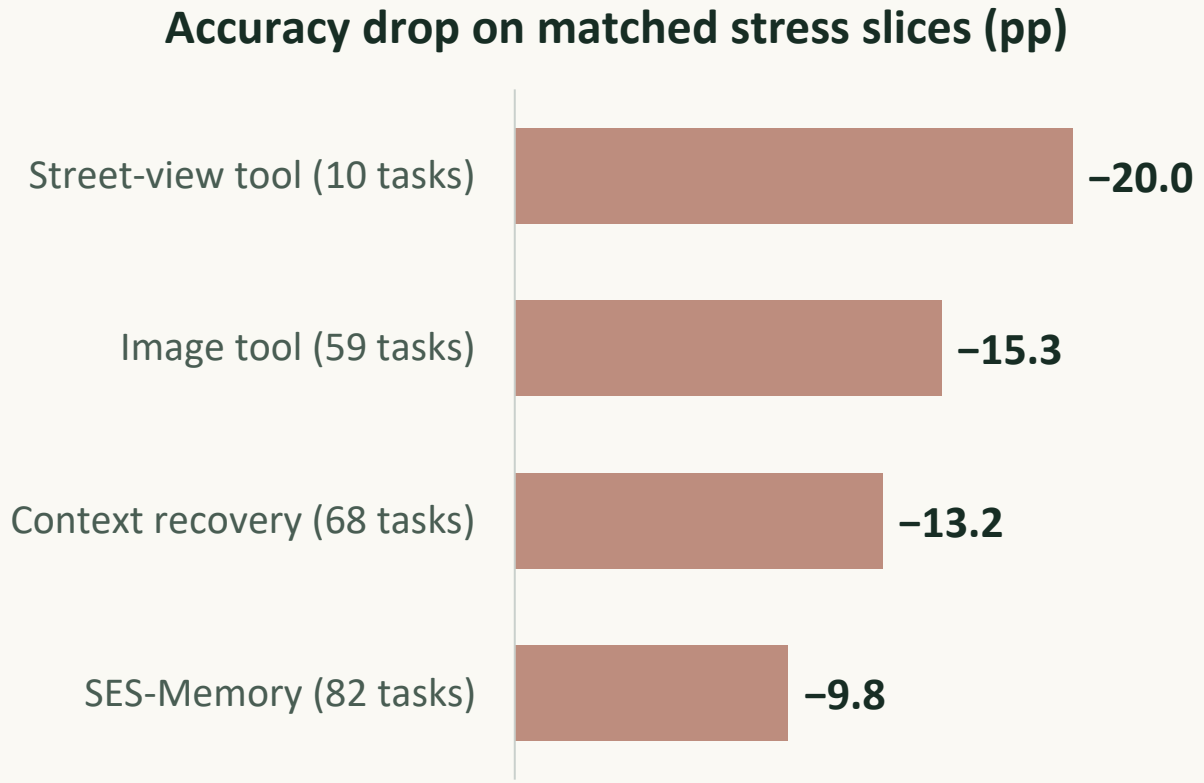
50–88% fewer tokens per task

Same GAIA tasks and GPT-5 controller and worker backbone

Co-Sight 11.8 min/task vs. Lemon 17.5 min/task

One attempt per task. Lemon also calls other hosted models.

Ablations and controller interventions



Each mechanism is tested on its own task slice.

Full GAIA benchmark

All four mechanisms removed

91.36% to 81.73%

−9.63 pp on 301 tasks

Controller prompt interventions

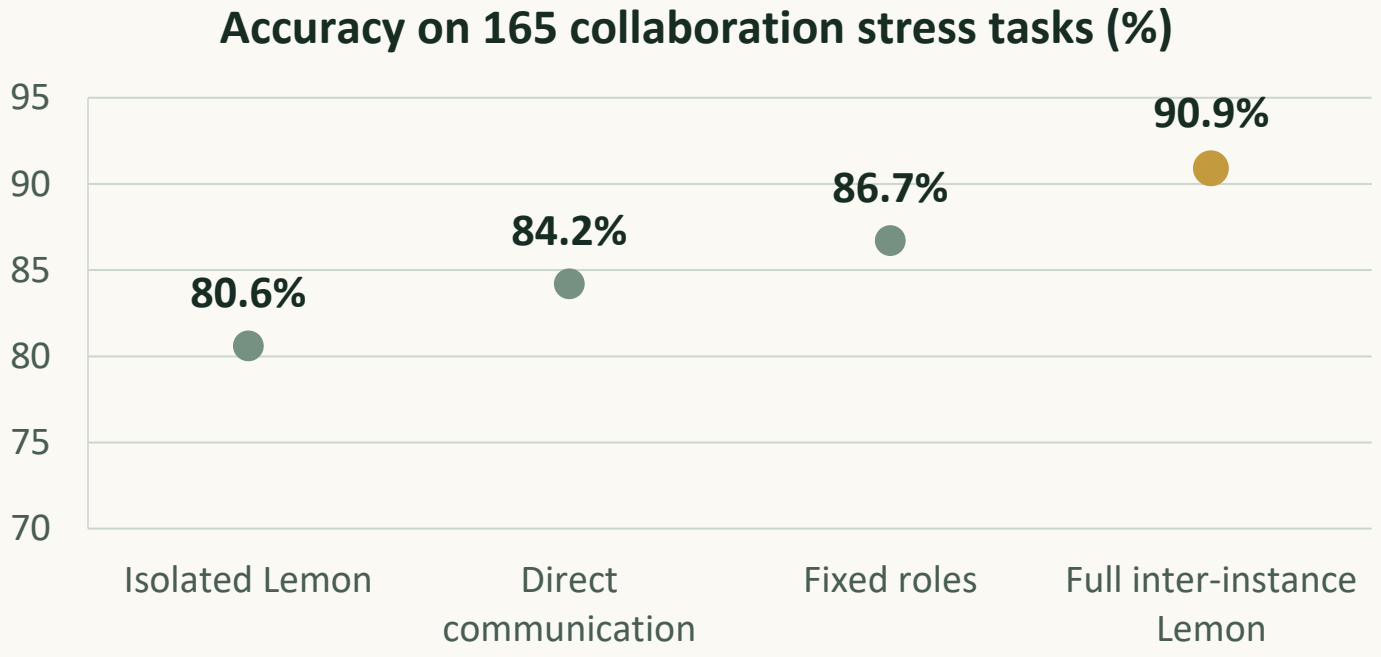
Same backbone, tools, and caps

Evidence state ignored	**−4.32 pp**
Evidence clauses removed	**−6.64 pp**
Placebo edit	**−0.33 pp**

One run per controller variant

Evidence exchange across agents

Expose expertise → Share proposals with evidence → Critique unsupported claims → **Claim-level consensus** → Reuse useful artifacts



Full vs. isolated Lemon

+10.3 pp

Accuracy gain

≈3.4×

Tokens per correct answer

1.08 vs. 0.32 MTok

The isolated agent differs in tools, memory, and total budget.

Evidence risk guides execution within and across agents

Within one agent

The current evidence state guides each decision.

Across agents

Agents exchange evidence and reach claim-level consensus.

Limitations and next steps

The controller uses a hand-written prompt.
The evaluation covers two benchmarks and one stress set.
Next steps include learned control, broader evaluation, and safer memory sharing.



Thank you