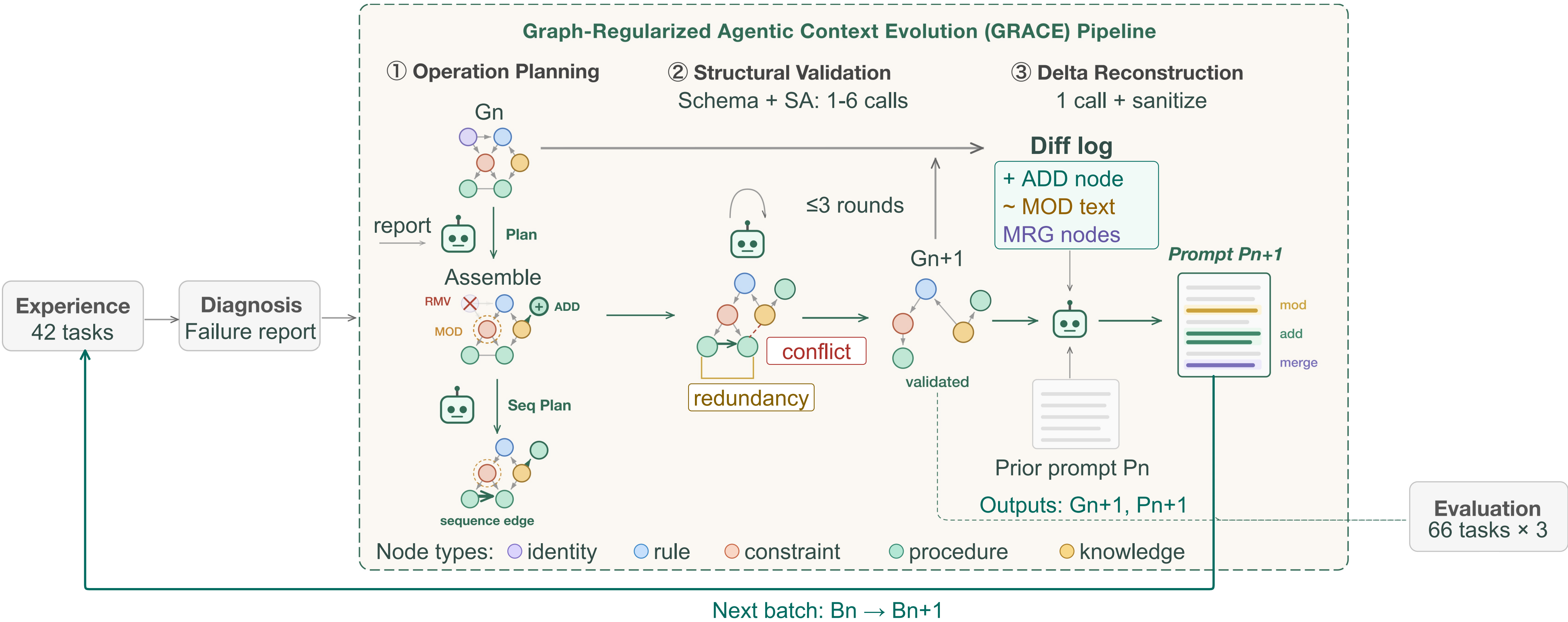


Scoped Verification for Reliable Lifelong Agentic Context Evolution under Distribution Shift

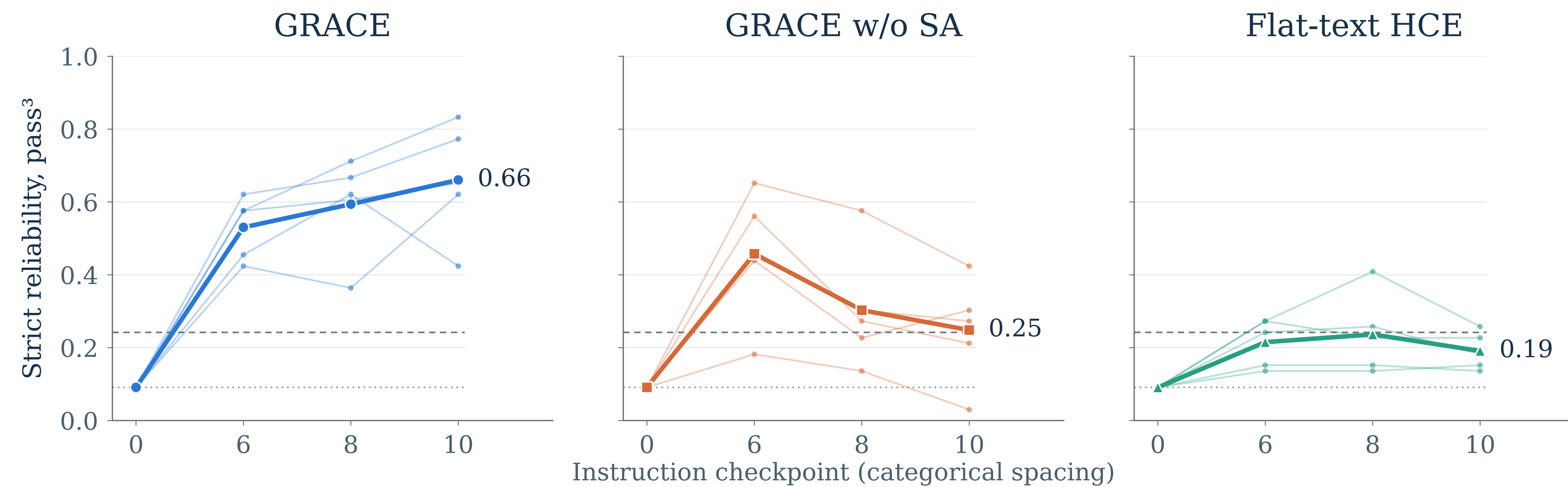
Dan C. Hsu^{1,2} · Luke Lu^{*1,3} ^{*} luke@redmindresearch.org
¹ RedMind Research Foundation ² National Taiwan University ³ Workato, Inc

Check local changes. Preserve accumulated knowledge.



Structural analysis (SA) uses LLM judgments to check the edited units' typed neighborhood for contradictions and redundancy, then repair conflicts and consolidate redundant content. Rule, constraint and procedure are informal norm categories; patches preserve untouched text.

Reliability improves across repeated revisions



STRICT RELIABILITY AT UPDATE 10

66.1% ± 15.8 pp
GRACE · mean ± SD

+47.0 percentage points vs flat-text HCE

HCE: 19.1% ± 5.1 pp · Without SA: 24.8% ± 14.4 pp

What the ablation changes

Same model, tools, harness and update schedule.

	GRACE	Without SA	Flat-text HCE
Substrate	Typed graph	Typed graph	Flat text
Validation	Schema + scoped SA	Schema only	In-call self-check
Write-back	Anchored patches	Anchored patches	Full re-emission
Calls / update	6.3	3.1	1.0

SA: structural analysis. Calls: mean over 4 runs with ledgers. Internal call budgets are not matched.

Controlled shifts, held-out tests

r²-bench telecom · Gemini 2.5 Flash

10 batches × 42 tasks · 5 runs / condition

Batches **A** **A** **B** **B** **A** **A** **B** **B** **A** **A**

A: MMS-heavy B: mobile-data-heavy

66 held-out tasks × 3 trials. pass³ requires all three trials to succeed; the test set never guides evolution.

TAKEAWAY

Typed graphs make growing instructions locally checkable and repairable.

Scope: one domain, model and harness. A budget-matched flat-text control remains open.