UNSW SYDNEY | Australia's Global University

Code

MAGE: Multi-Agent Self-Evolution with Co-Evolutionary Knowledge Graphs

Ruiyi Yang[1], Zechen Li[1], Hao Xue[2,1], Imran Razzak[3,1], Flora D. Salim[1]

[1] University of New South Wales
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] Mohamed Bin Zayed University of Artificial Intelligence

Background

Parameter updates

Store learning in weights, but require a model that can be updated.

Natural-language feedback

Can become generic and lose the structure of recurring failures.

Flat episodic memory

Preserves experience but exposes little about skill dependencies or task coverage.

Goal

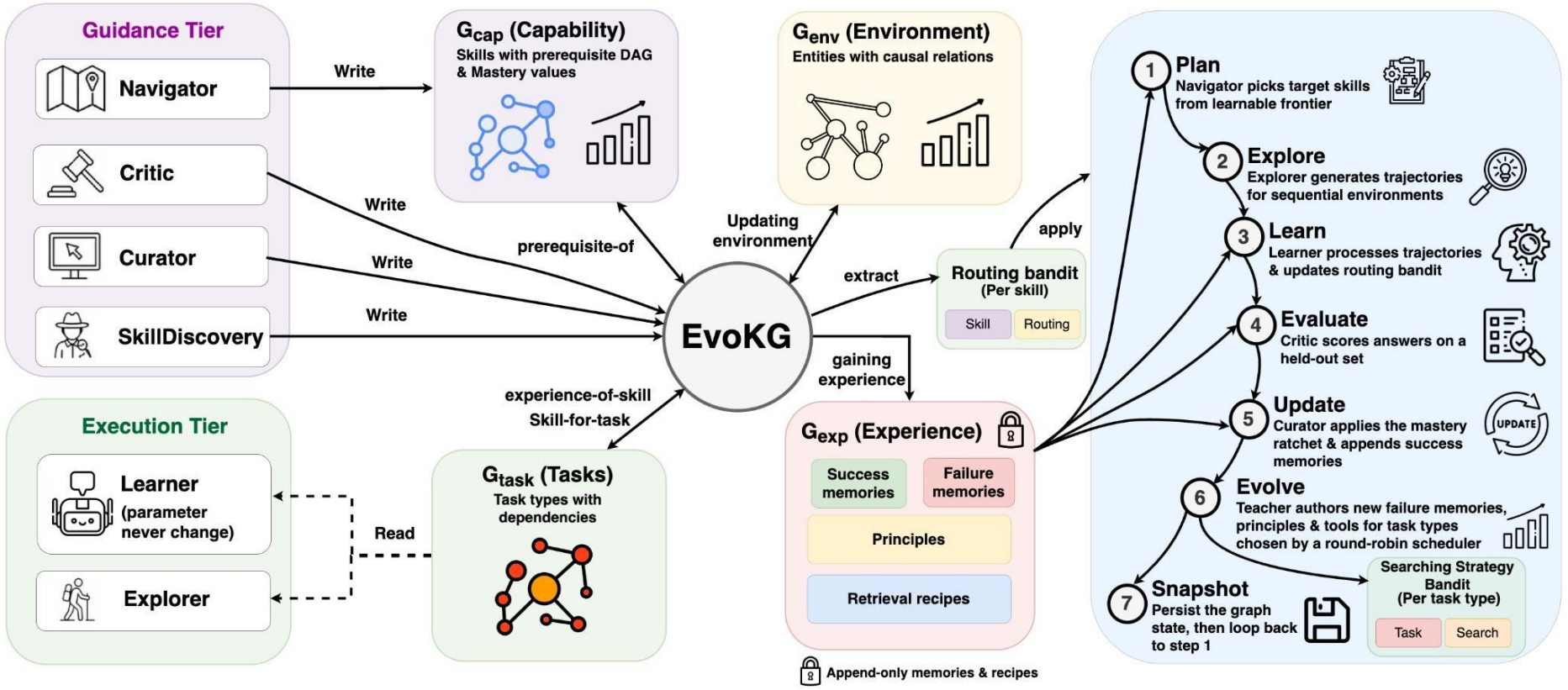
Can a frozen execution model improve by evolving its external knowledge?

EvoKG carries experience across iterations

Link capabilities, task types, experience and environment context.

Correct learner answers become future demonstrations.
Teacher-written corrections preserve lessons from failures.

MAGE framework



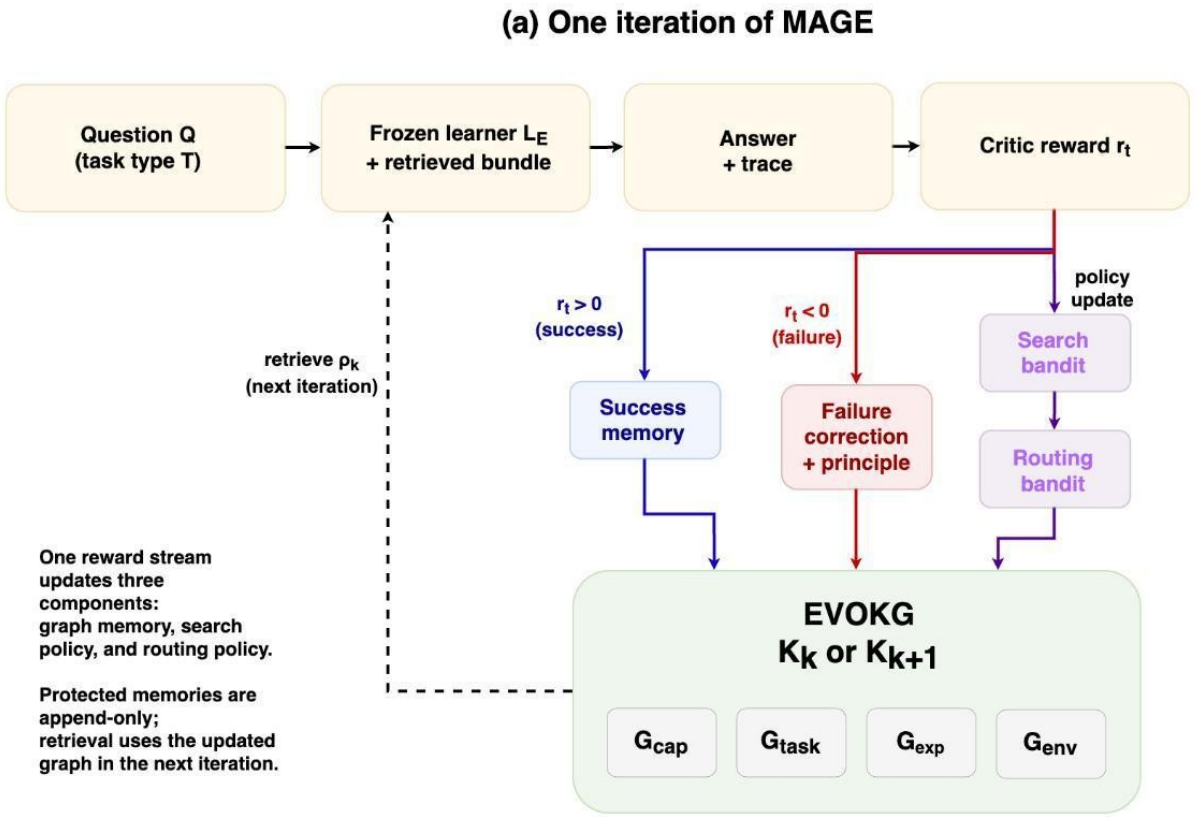
Guidance writes the graph during evolution. The frozen execution model reads it at inference.

EvoKG: four connected subgraphs

Subgraph	Stored state	Role
Capability	Skills, mastery, prerequisites	Dependency-aware execution
Task	Task types and dependencies	Curriculum and memory filtering
Experience	Successes, corrections, principles	Reusable inference guidance
Environment	Observed entities and context	Sequential interaction support

Protected success, failure and principle memories are append-only.

One iteration of MAGE



One correctness reward

Store correct traces as success memories.

Write corrections for errors.

Update the graph and active search / routing policies.

Rollback protects mutable state.
Protected memories remain.

Success and failure memories

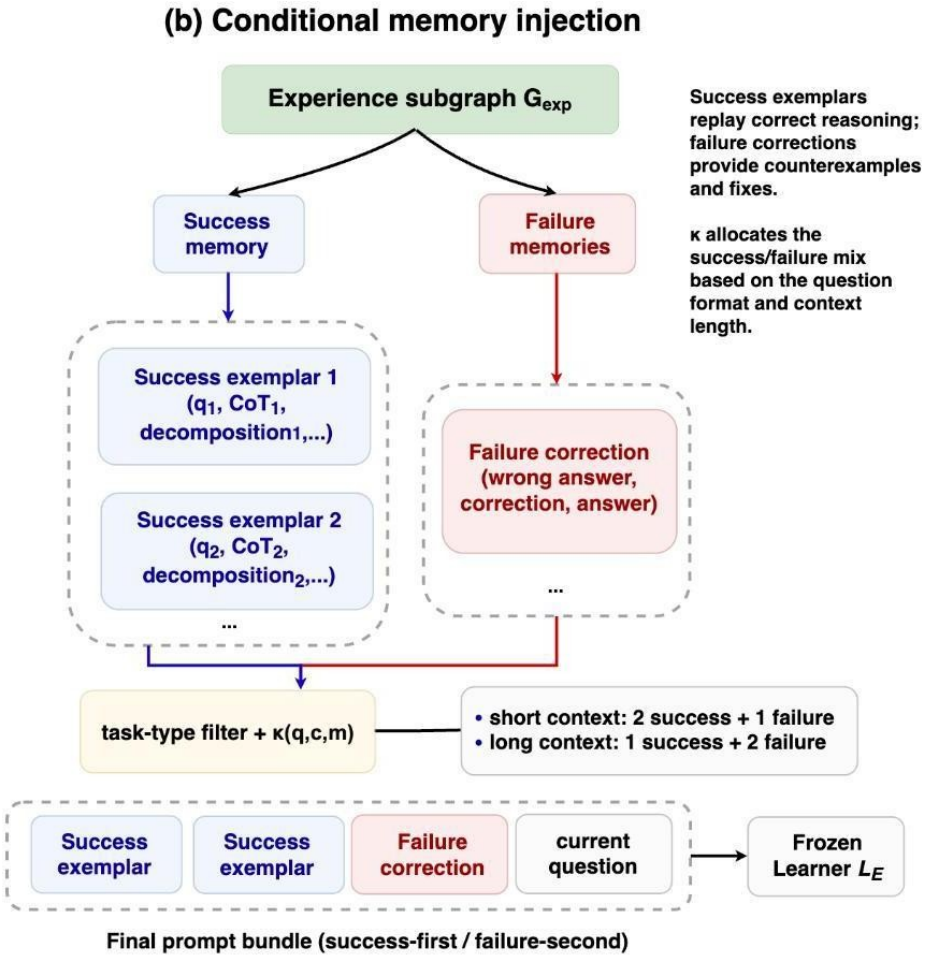
Success memories

The learner's own correct traces preserve reusable reasoning steps.

Failure memories

Guidance-written corrections explain observed mistakes.

Retrieval filters by task type and adapts the mixture to query context.



Structural support for evolution

Design rule	Structural property
Append-only protected memory	The retained information set does not shrink
Recency-weighted task selection	A bounded gap between visits to observed task types
Asymmetric mastery update	Bounded decay after a poor measurement
Task-filtered retrieval	Retrieval-side support under an idealized error model

These results do not guarantee monotonic LLM accuracy.

UNSW SYDNEY

Evaluation protocol

Task family	Benchmarks
Math	GSM8K, RealMath
Question answering	HotpotQA, WebQA
Spatio-temporal	STBench
Finance / medicine	FinQA, MedQA-USMLE
Sequential interaction	Crafter, WebShop

Frozen execution

Qwen3-8B
Llama-3.1-8B-Instruct

Separate evolution and test pools

Freeze the graph and bandits before reporting test results.
No test example enters memory.

Core suite: 200 held-out questions per benchmark. Main MAGE cells: five seeds, mean ± SD.

Main reasoning results

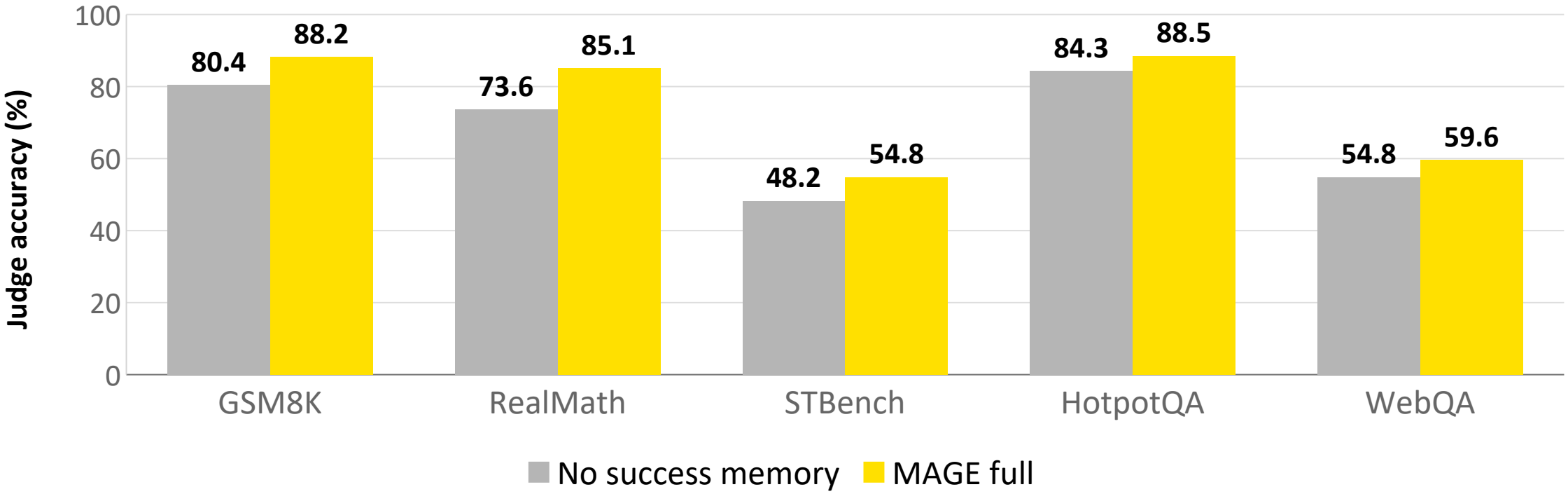
Benchmark	Best prompt baseline	MAGE Sonnet	MAGE Opus	Best gain (points)
GSM8K	82.5	88.2 ± 2.9	90.4 ± 3.8	+7.9
HotpotQA	84.5	88.5 ± 2.0	91.5 ± 1.8	+7.0
WebQA	55.0	59.6 ± 3.4	62.7 ± 1.8	+7.7
STBench	37.0	54.8 ± 2.1	57.8 ± 2.9	+20.8
RealMath	53.5	85.1 ± 6.7	82.8 ± 3.5	+31.6

Largest gains: RealMath and STBench

Table 1. Frozen Qwen3-8B. Semantic-judge accuracy (%), 200 questions per benchmark. MAGE: n = 5.

UNSW SYDNEY

Ablation: success memory



Success memories add +7.0 points on average across the five tasks.

Table 5. No-success keeps teacher-written failure memories. Means shown; full mean ± SD table in notes.

UNSW SYDNEY

Domain and sequential results

Llama-3.1-8B domain extensions

Benchmark	0-shot CoT	MAGE Sonnet	MAGE Opus
FinQA (numeric EM)	58.0	67.5 ± 1.7	69.0 ± 1.5
MedQA (letter EM)	64.5	74.6 ± 2.3	79.8 ± 1.9

Environment / metric	Reference	MAGE Sonnet	MAGE Opus
Crafter peak unlock (%)	25.5 ± 3.2	33.5 ± 2.3	37.9 ± 3.5
WebShop task score (%)	87.6	90.2 ± 2.3	89.7 ± 1.6
WebShop strict success (%)	78.9	84.0 ± 1.8	82.0 ± 1.0

Tables 2–3. MAGE: n = 5. Crafter reference: Llama-3.1-8B. WebShop reference: published RetroAgent IC.

UNSW SYDNEY

Inference cost

0 guidance-model calls at inference

Benchmark	0-shot CoT	SC (N = 10)	MAGE
GSM8K	4.98 s	52.80 s	5.66 s
HotpotQA	2.82 s	30.74 s	11.64 s
WebQA	4.41 s	45.33 s	8.83 s
STBench	6.43 s	51.10 s	28.33 s
RealMath	5.09 s	42.30 s	18.83 s

Appendix N, Table 15. Seconds per question on one H100 80GB. Graph construction cost is separate.

UNSW SYDNEY

Memory and execution controls

Benchmark	Control	MAGE EM / F1	Control EM / F1
WebQA	A-MEM	62.7 / 67.1	62.7 / 69.4
STBench	A-MEM	57.2 / 61.3	51.7 / 52.0
RealMath	A-MEM	82.4 / 84.5	69.7 / 66.8
RealMath	No-chain	82.4 / 84.5	72.0 / 74.0
HotpotQA	MemSkill	54.0 / 71.67	40.5 / 55.34

A-MEM has higher WebQA F1. Prerequisite traversal adds compute.

Table 6. Task-native EM/F1 (%). RealMath: 600 paired runs. MemSkill: matched 1,024-token answer budget.

UNSW SYDNEY

Limitations and future work

Recurring structure matters

Benefits depend on reusable task patterns and the quality of graph writes.

Component benefits depend on activation

Skill routing helps one activated setting. Task search has no significant measured gain.

Long-term memory management remains open

Known supersession can be filtered, but discovering new contradictions remains unresolved.

Thanks for listening

Code

Frozen models can reuse experience stored in EvoKG

Ruiyi Yang

ruiyi.yang@unsw.edu.au

github.com/RuiyiYang01/mage-alg

Backup: prompting baselines

Benchmark	0-shot	8-shot	SC10	ReAct	Reflexion	Sonnet	Opus
GSM8K	64.5	82.5	72.0	80.1	63.0	88.2	90.4
HotpotQA	84.5	78.5	80.5	70.5	78.5	88.5	91.5
WebQA	55.0	42.0	46.5	35.5	34.5	59.6	62.7
STBench	35.5	12.0	20.0	31.8	37.0	54.8	57.8
RealMath	53.5	32.5	33.5	43.5	34.5	85.1	82.8

Table 1. Qwen3-8B semantic-judge accuracy (%). MAGE means shown; SDs appear on the main results slide.

Backup: local teachers and transfer

Local Qwen3-32B guidance on RealMath

Control / execution setting	Gain in EM	Gain in F1	EM 95% CI
No traversal / Qwen3-8B	+5.8	+8.4	[3.3, 8.3]
Standalone / Qwen3-8B	+12.0	+16.2	[9.2, 14.8]
Standalone / Llama transfer	+9.5	+13.5	[3.5, 15.5]

Three independently evolved graphs. Qwen3 controls pool 600 questions.
The Llama audit swaps only the learner in a frozen locally evolved graph.

Backup: component attribution

Component / audit	Evidence and boundary
Skill routing	Activated STBench: +4.0 EM, 95% CI [1.5, 7.0]
Task search	WebQA: +1.0 overall, CI [-2.5, 4.5]
Task filtering	Higher in-task precision, no significant accuracy gain
Rollback	43 harmful-update interventions, observational audit
Memory lifecycle	Known obsolete records can be excluded from retrieval

Appendices S and U. Null and conditional results limit claims about the necessity of each component.

Backup: retrieved memory example

Tom has 3 times as many marbles as Sue.
Sue has half as many as Anne. Anne has 24.
How many marbles do they have in total?

Retrieved guidance

Success traces preserve multi-step arithmetic. A correction memory reminds the learner to sum every quantity after applying the ratios.

Anne = 24 Sue = 12 Tom = 36 Total = 72