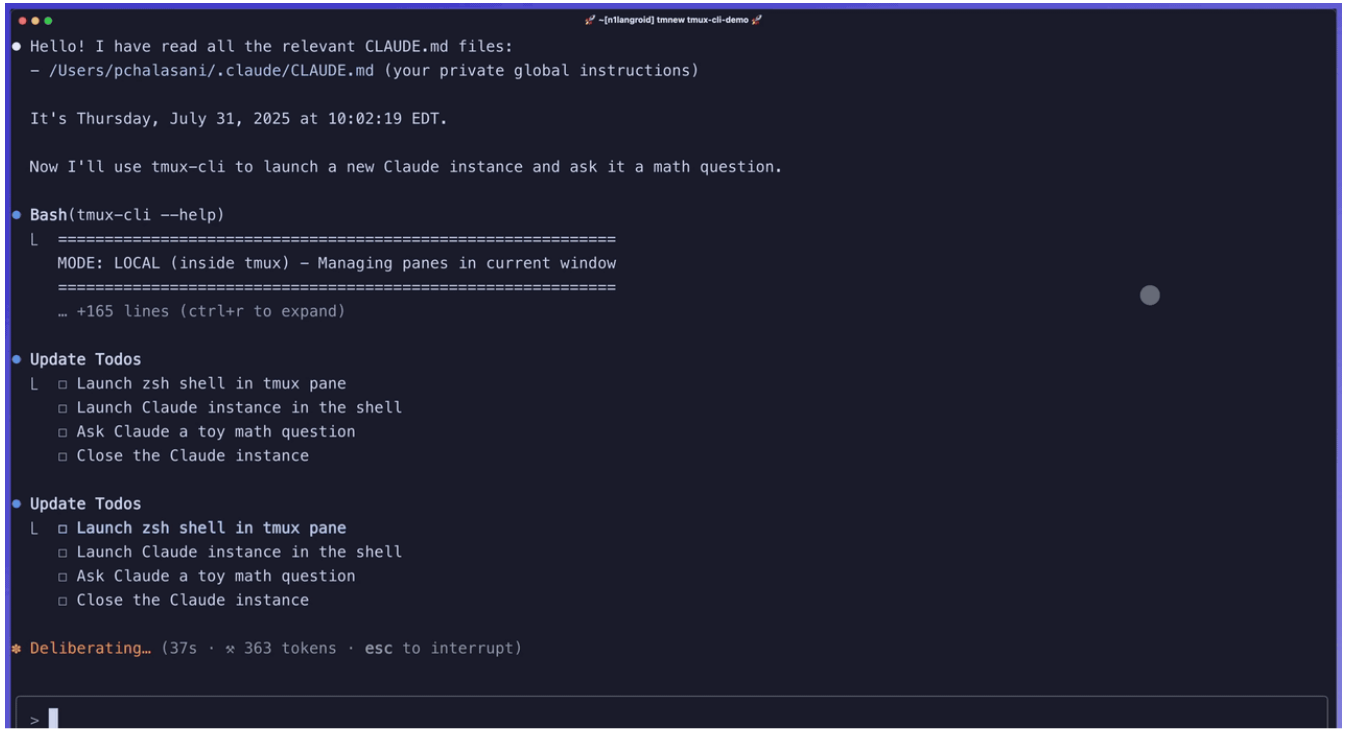
Self-Compacting Language Agents

Daniel Khashabi

Just another day in the office …

```
🚀 ~[n1langroid] tmnew tmux-cli-demo 🚀

● Hello! I have read all the relevant CLAUDE.md files:
  - /Users/pchalasani/.claude/CLAUDE.md (your private global instructions)

  It's Thursday, July 31, 2025 at 10:02:19 EDT.

  Now I'll use tmux-cli to launch a new Claude instance and ask it a math question.

● Bash(tmux-cli --help)
  └  ============================================================
     MODE: LOCAL (inside tmux) - Managing panes in current window
     ============================================================
     … +165 lines (ctrl+r to expand)

● Update Todos
  └  ☐ Launch zsh shell in tmux pane
     ☐ Launch Claude instance in the shell
     ☐ Ask Claude a toy math question
     ☐ Close the Claude instance

● Update Todos
  └  ☐ Launch zsh shell in tmux pane
     ☐ Launch Claude instance in the shell
     ☐ Ask Claude a toy math question
     ☐ Close the Claude instance

✻ Deliberating… (37s · ⚒ 363 tokens · esc to interrupt)

> 

  ? for shortcuts                ✗ Auto-update failed · Try claude doctor or npm i -g @anthropic-ai/claud

tmux-cli-demo  1:node
```

Just another day in the office …



Heads-up: I'm essentially at my context limit, so a /compact before the next step would keep things clean.

Just another day in the office …

- Untimely compaction
- The agent forget details midway debugging!!

```
Heads-up: I'm essentially at my context limit, so a /compact before the next step would keep things clean.
```

“Compaction” — defining it

“Compaction” — defining it

- Compaction :=
  1. condenses that context into a short summary,
  2. releases the memory, and
  3. generation continues from the summary.

- Agents populate a single rolling context
  - Thoughts, tool calls, observations, etc.
  - Some useful and some junk.

“Compaction” — defining it

- Compaction :=
  1. condenses that context into a short summary,
  2. releases the memory, and
  3. generation continues from the summary.

- Agents populate a single rolling context
  - Thoughts, tool calls, observations, etc.
  - Some useful and some junk.

The challenge: context rot

- Increasing context lowers attention fidelity

The challenge: context rot

- Increasing context lowers attention fidelity

The effective context window of models is much smaller than what they're claimed to be.

Models	Claimed Length	Effective Length
Llama2 (7B)	4K	-
Gemini-1.5-Pro	1M	>128K
GPT-4	128K	64K
Llama3.1 (70B)	128K	64K
Qwen2 (72B)	128K	32K
Command-R-plus (104B)	128K	32K
GLM4 (9B)	1M	64K
Llama3.1 (8B)	128K	32K
GradientAI/Llama3 (70B)	1M	16K
Mixtral-8x22B (39B/141B)	64K	32K
Yi (34B)	200K	32K
Phi3-medium (14B)	128K	32K
Mistral-v0.2 (7B)	32K	16K

Hsieh et al. RULER: What's the Real Context Size of Your Long-Context Language Models? COLM, 2024.

The challenge: limited context length

How systems handle it today

- Nowadays compaction is used in all agentic tools.

- It used to be that it triggered near the point of exceeding the window.
- Increasingly compaction is frequent and invisible.

- It works more or less, **but**:
  - Hard to study it scientifically.
  - Not perfect and not doing it at the right time will be harmful.

How systems handle it today

- Nowadays compaction is used in all agentic tools.

- It used to be that it triggered near the point of exceeding the window.
- Increasingly compaction is frequent and invisible.

- It works more or less, **but**:
  - Hard to study it scientifically.
  - Not perfect and not doing it at the right time will be harmful.

How systems handle it today

- Nowadays compaction is used in all agentic tools.

- It used to be that it triggered near the point of exceeding the window.
- Increasingly compaction is frequent and invisible.

- It works more or less, **but**:
  - Hard to study it scientifically.
  - Not perfect and not doing it at the right time will be harmful.

How systems handle it today

- Nowadays compaction is used in all agentic tools.

- It used to be that it triggered near the point of exceeding the window.
- Increasingly compaction is frequent and invisible.

- It works more or less, **but**:
  - Hard to study it scientifically.
  - Not perfect and not doing it at the right time will be harmful.

Research question

Research question

Can LLM agents *off-the-shelf* recognize their own context rot and compact accordingly?

Research question

Can LLM agents *off-the-shelf* recognize their own context rot and compact accordingly?

How?

Research question

Can LLM agents *off-the-shelf* recognize their own context rot and compact accordingly?

When?

How?

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

P

QUESTION (BrowseComp)

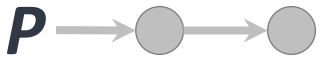
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

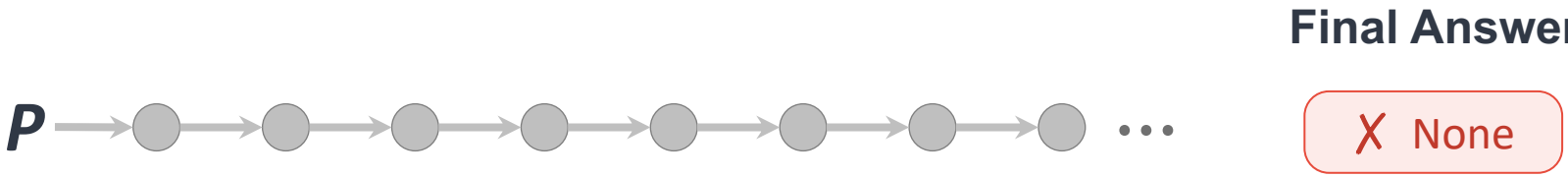
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

No Compression baseline

Final Answer:

✗ None

Run out of context runway + costly.

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Fixed-interval compaction (every *k* turns or % of context limit)

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

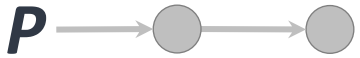
Fixed-interval compaction (every *k* turns or % of context limit)

P

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

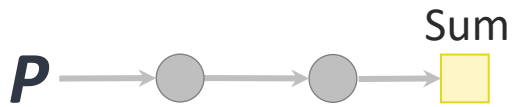
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

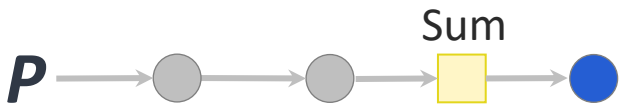
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

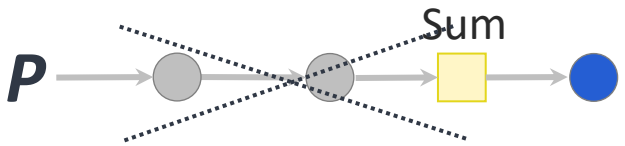
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

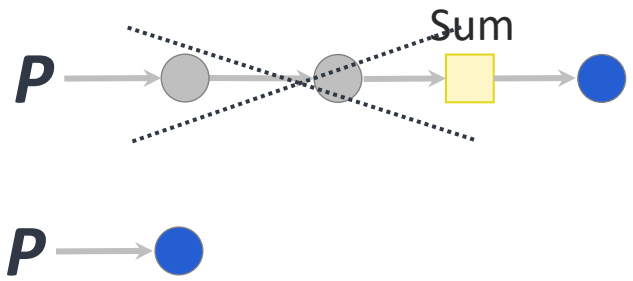
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

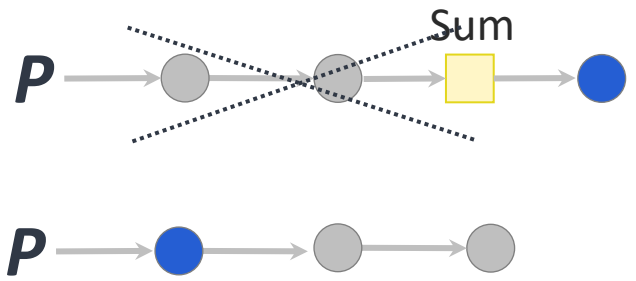
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

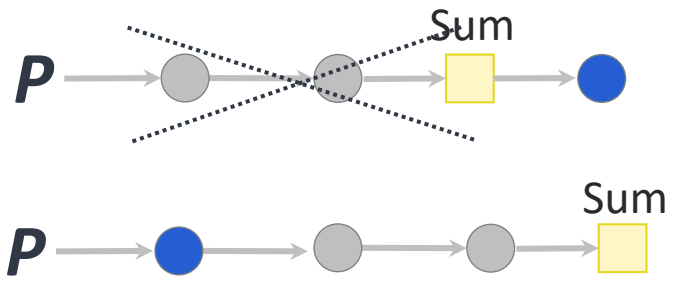
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

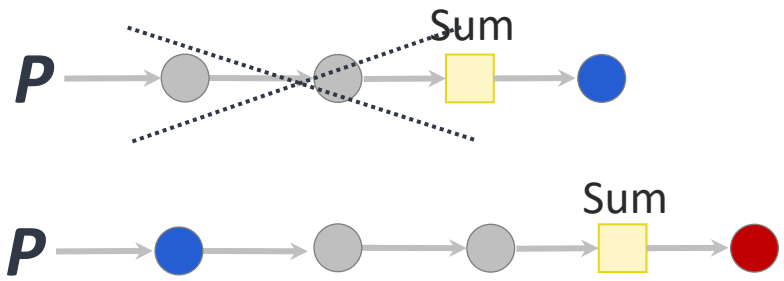
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

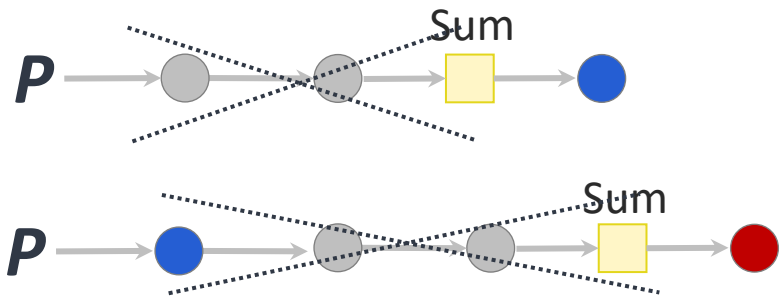
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

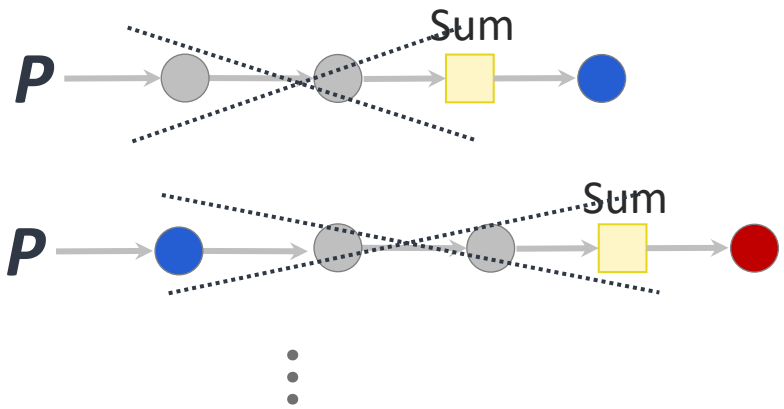
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

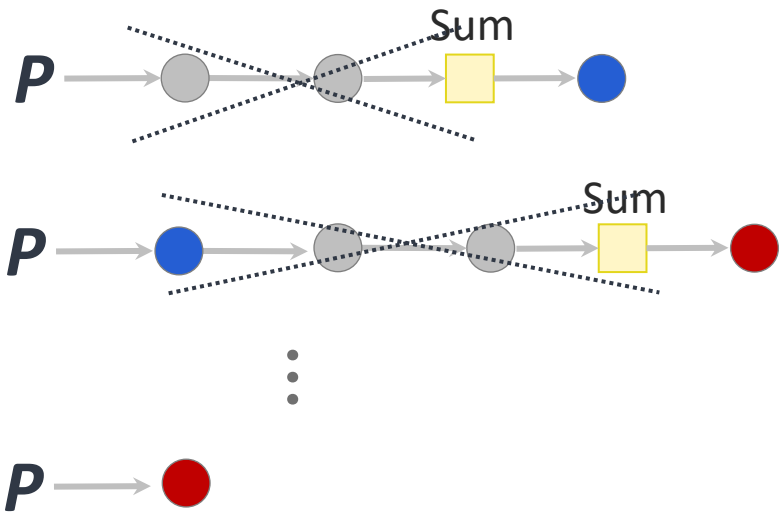
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

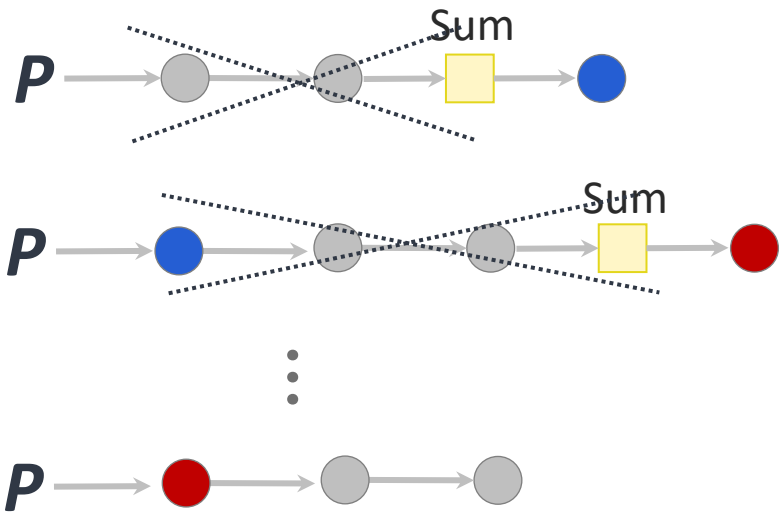
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

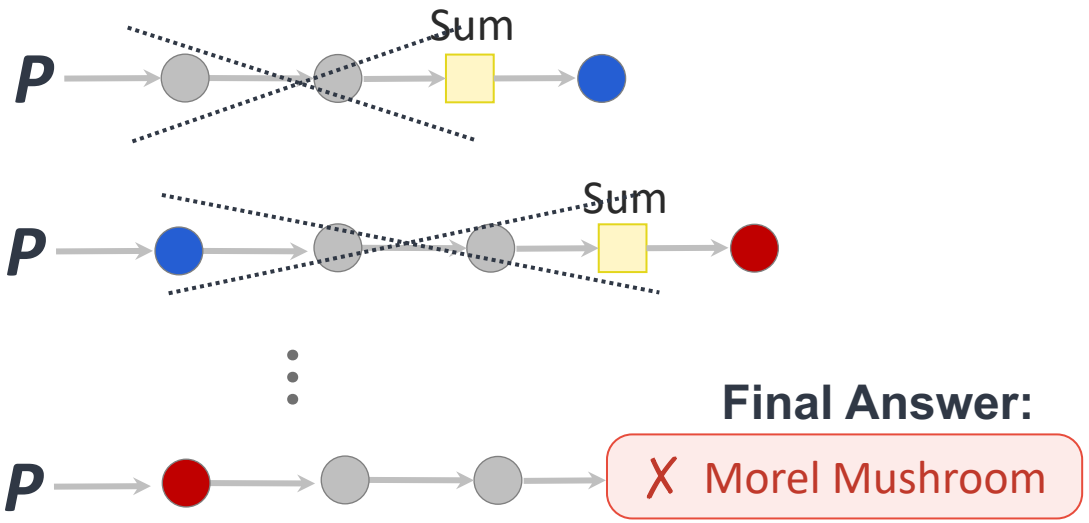
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

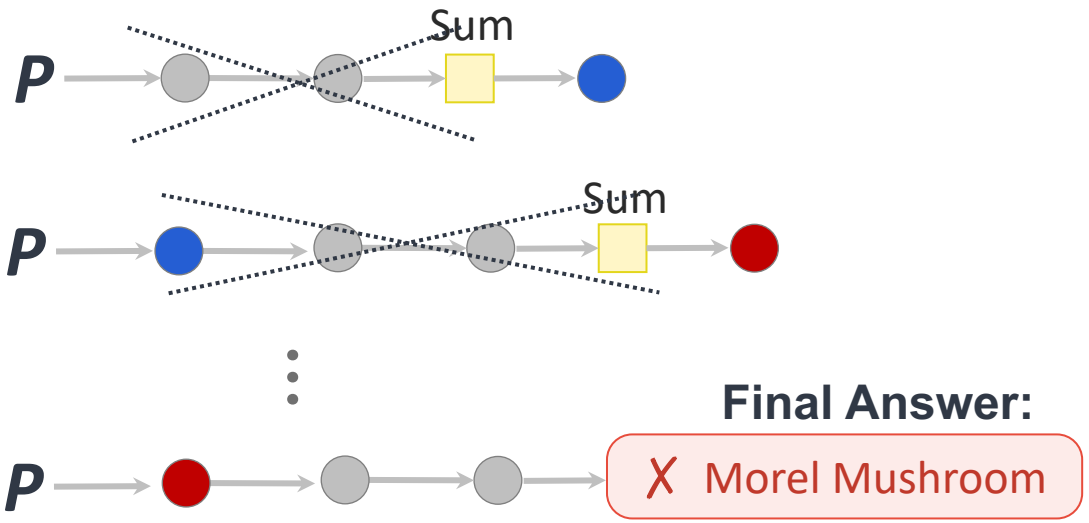
Fixed-interval compaction (every *k* turns or % of context limit)



QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Fixed-interval compaction (every *k* turns or % of context limit)

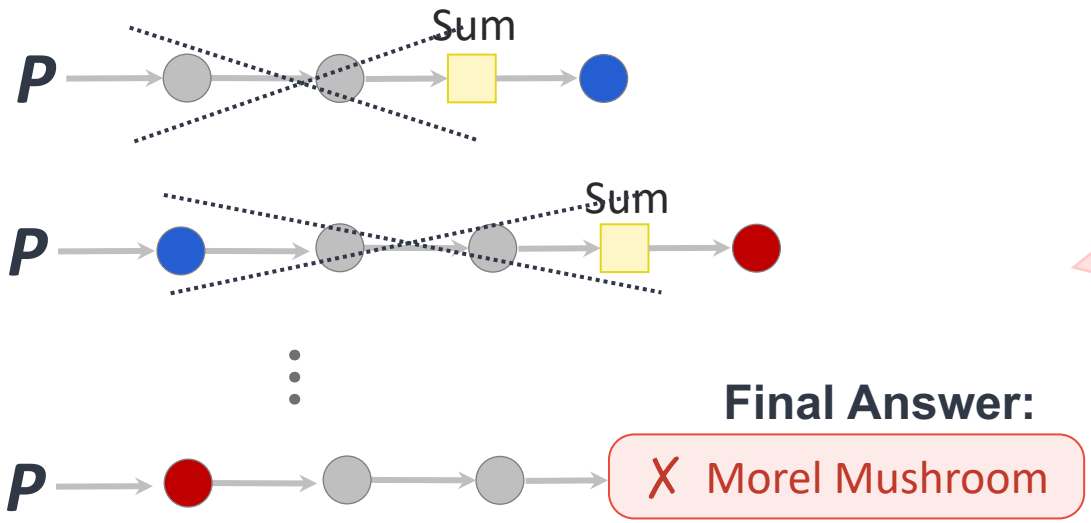


AgentFold, 2025; FoldAct, 2025, Context-Folding, 2025; Reasoning Cache, 2026; InftyThink, 2026; IterResearch, 2026; ReSuM, 2026;…

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Fixed-interval compaction (every *k* turns or % of context limit)

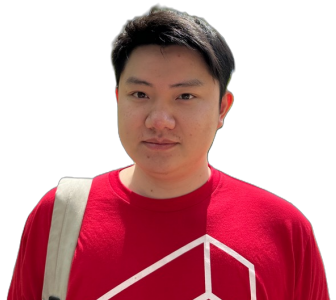


poorly timed summarization — drops verified facts; restates intent vaguely.

Self-Compact

Self-Compacting Language Model Agents

Tianjian Li♠ Jingyu Zhang♠ William Jurayj♠ Xi Wang♠ Chuanyang Jin♠
Mehrdad Farajtabar♡ Eric Nalisnick♠ Daniel Khashabi♠



Self-Compact

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each turn, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each turn, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each turn, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each turn, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each turn, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

P

Self-Compact, visually

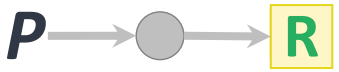
QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

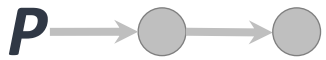
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

$\in$ {COMPACT, **CONTINUE**}

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

$\in$ {COMPACT, CONTINUE}

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

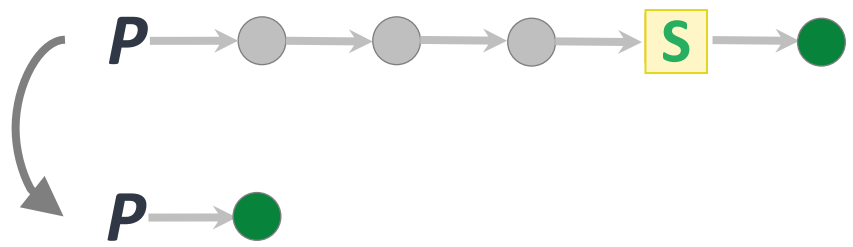
QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

Self-Compact, visually

QUESTION (BrowseComp)

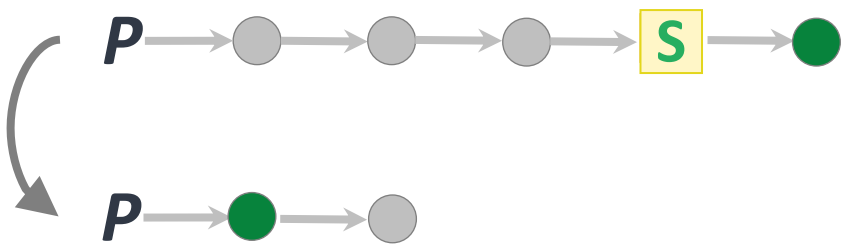
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Self-Compact, visually

QUESTION (BrowseComp)

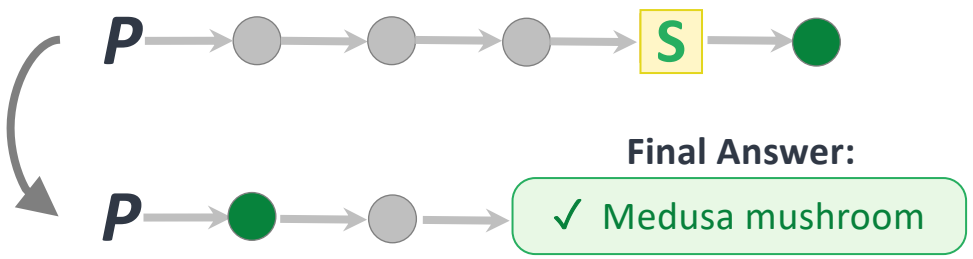
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Self-Compact, visually

QUESTION (BrowseComp)

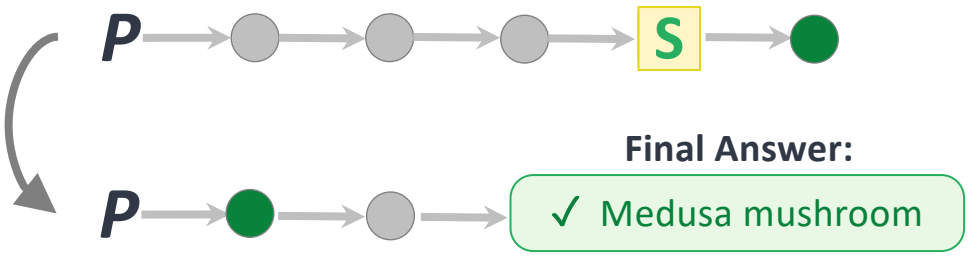
Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.

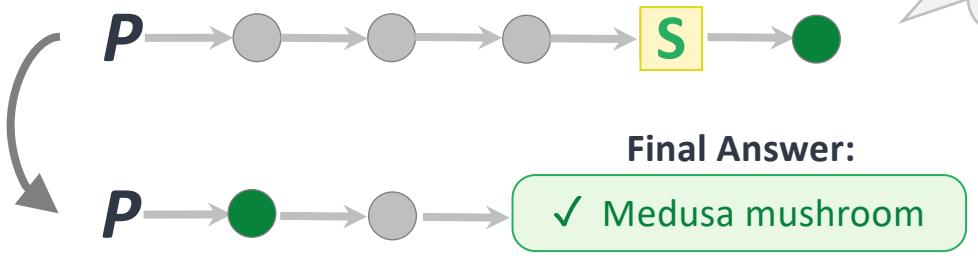


adaptive summarization — consolidates facts; lets model escape from loops.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



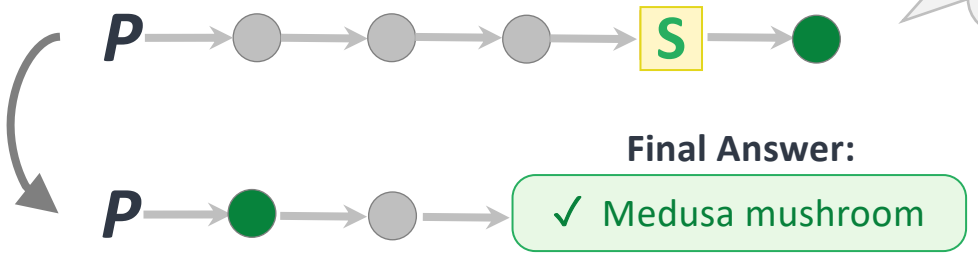
Minimal overhead because of KV-reuse

adaptive summarization — consolidates facts; lets model escape from loops.

Self-Compact, visually

QUESTION (BrowseComp)

Input prompt (*P*): Identify a rare fungus appearing in clusters after rainfall, with raised scales on its cap, named by a French expert in the 1980s, with potential antifungal properties. Its English name matches a 1980s film character; the film was inspired by a 1970s bronze statuette. Two-word name; first word has 3 syllables, ends in vowel.



Minimal overhead because of KV-reuse

Shortens a long context (less tokens to attend)

adaptive summarization — consolidates facts; lets model escape from loops.

Self-Compact: Cost

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each step, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact: Cost

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each step, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Shortens a long context
(less tokens to attend)

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact: Cost

1. **Compaction tool (C):** how to compact (summarize)
2. **Rubric (R):** when to fire

- Run reasoning as usual.
- After each step, run **R** ∈ {COMPACT, CONTINUE}.
- If CONTINUE, continue thinking (remove rubric from history).
- If COMPACT, call **C** and replace the context w/ summary.

Minimal overhead because of KV-reuse

Shortens a long context (less tokens to attend)

Li et al. Self-Compacting Language Model Agents, 2026.

Self-Compact: Making it work

- The Rubric prompt is the crucial design component.
- Guidelines for when to compact we found effectives:

1. The model finished a subtask
   - is irrelevant to subsequent turn, and it only needs the result of the subtask (e.g. the model derived an intermediate result in math, the model verified a search result)
2. The model is stuck
   - compaction would give the model an opportunity to break out

Self-Compact: Making it work

- The Rubric prompt is the crucial design component.
- Guidelines for when to compact we found effectives:

1. The model finished a subtask
   - is irrelevant to subsequent turn, and it only needs the result of the subtask (e.g. the model derived an intermediate result in math, the model verified a search result)
2. The model is stuck
   - compaction would give the model an opportunity to break out

Self-Compact: Making it work

- The Rubric prompt is the crucial design component.
- Guidelines for when to compact we found effectives:

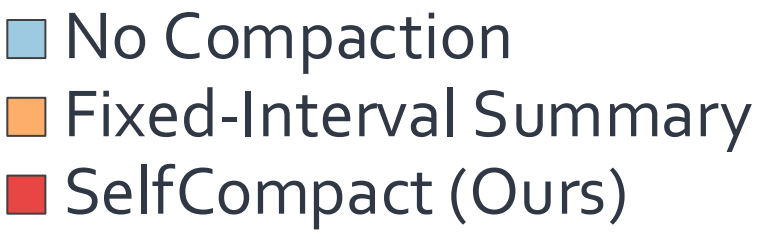
1. The model finished a subtask
   - is irrelevant to subsequent turn, and it only needs the result of the subtask (e.g. the model derived an intermediate result in math, the model verified a search result)
2. The model is stuck
   - compaction would give the model an opportunity to break out

Self-Compact: Making it work

- The Rubric prompt is the crucial design component.
- Guidelines for when to compact we found effectives:

1. The model finished a subtask
   - is irrelevant to subsequent turn, and it only needs the result of the subtask (e.g. the model derived an intermediate result in math, the model verified a search result)
2. The model is stuck
   - compaction would give the model an opportunity to break out

Empirical Results: Agentic Search



* Uses scaffold of Lee et al. 2026. “*Agentic Aggregation for Parallel Scaling of Long-Horizon Agentic Tasks*”

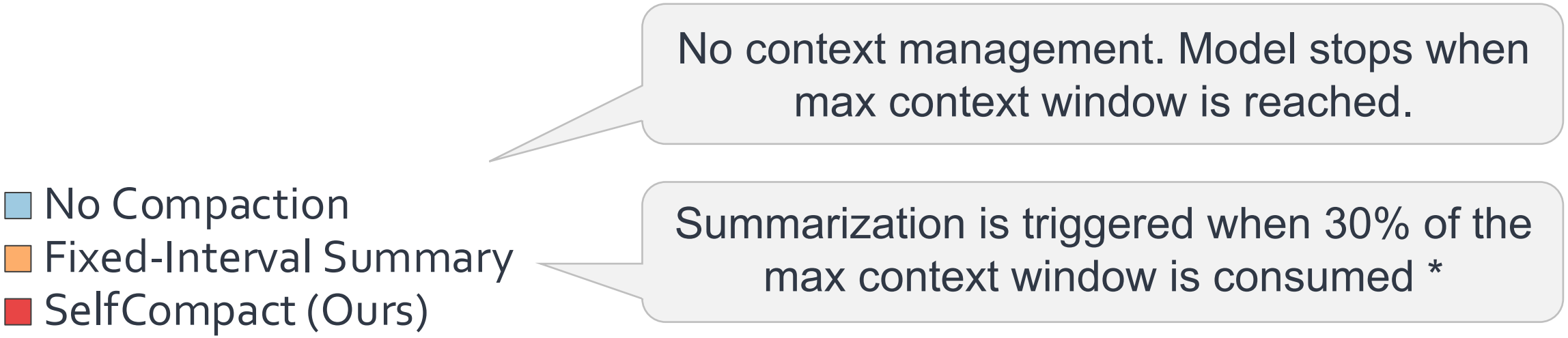
Empirical Results: Agentic Search

No context management. Model stops when max context window is reached.

- No Compaction
- Fixed-Interval Summary
- SelfCompact (Ours)

* Uses scaffold of Lee et al. 2026. “*Agentic Aggregation for Parallel Scaling of Long-Horizon Agentic Tasks*”

Empirical Results: Agentic Search



* Uses scaffold of Lee et al. 2026. "*Agentic Aggregation for Parallel Scaling of Long-Horizon Agentic Tasks*"

Empirical Results: Agentic Search

No context management. Model stops when max context window is reached.

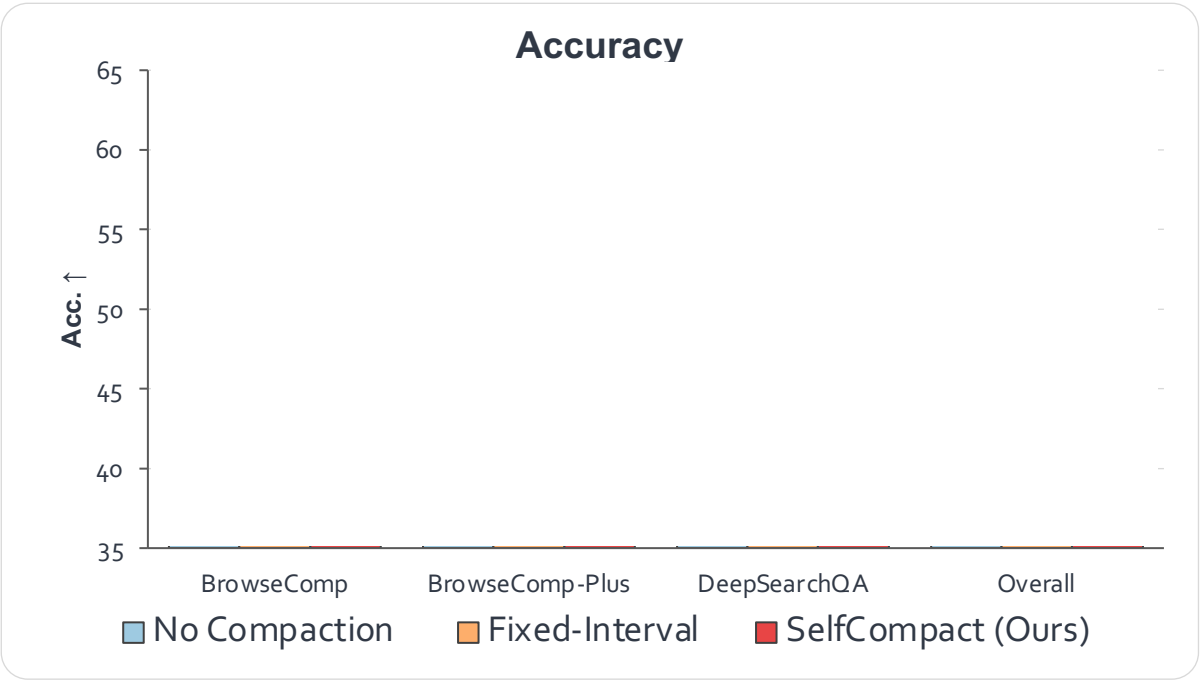
Summarization is triggered when 30% of the max context window is consumed *

Let the model decide when to compact.

* Uses scaffold of Lee et al. 2026. “*Agentic Aggregation for Parallel Scaling of Long-Horizon Agentic Tasks*”

Empirical Results: Agentic Search

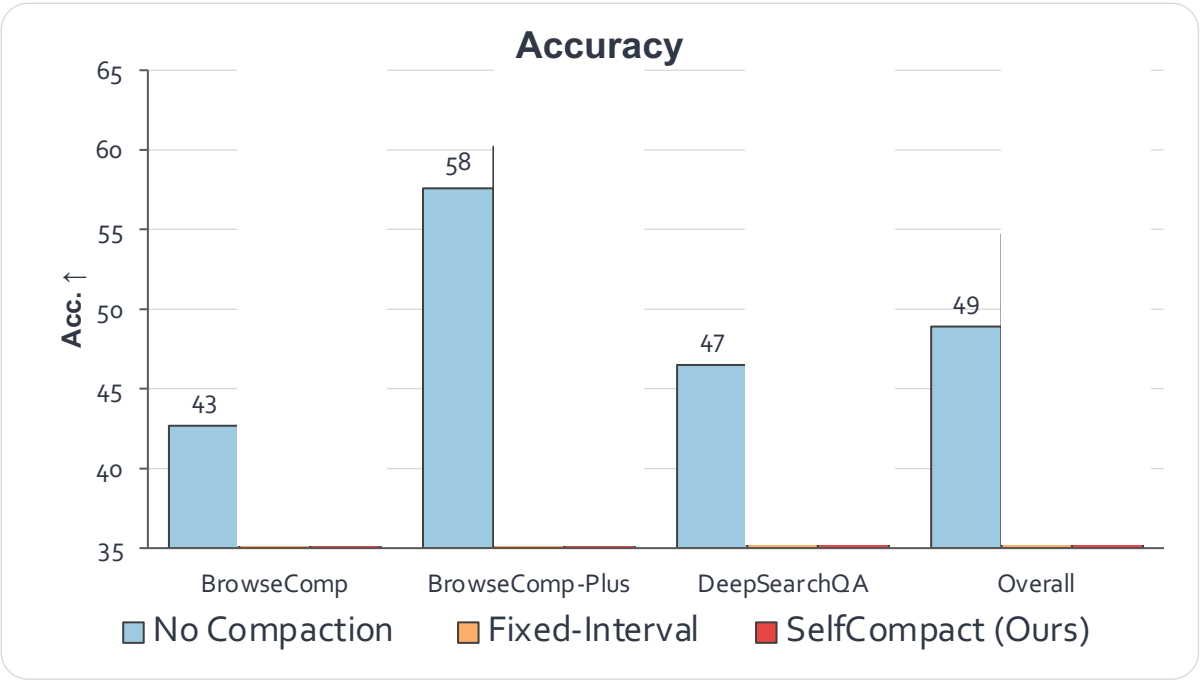
- Mimo-V2-Flash (309B; 15B active)

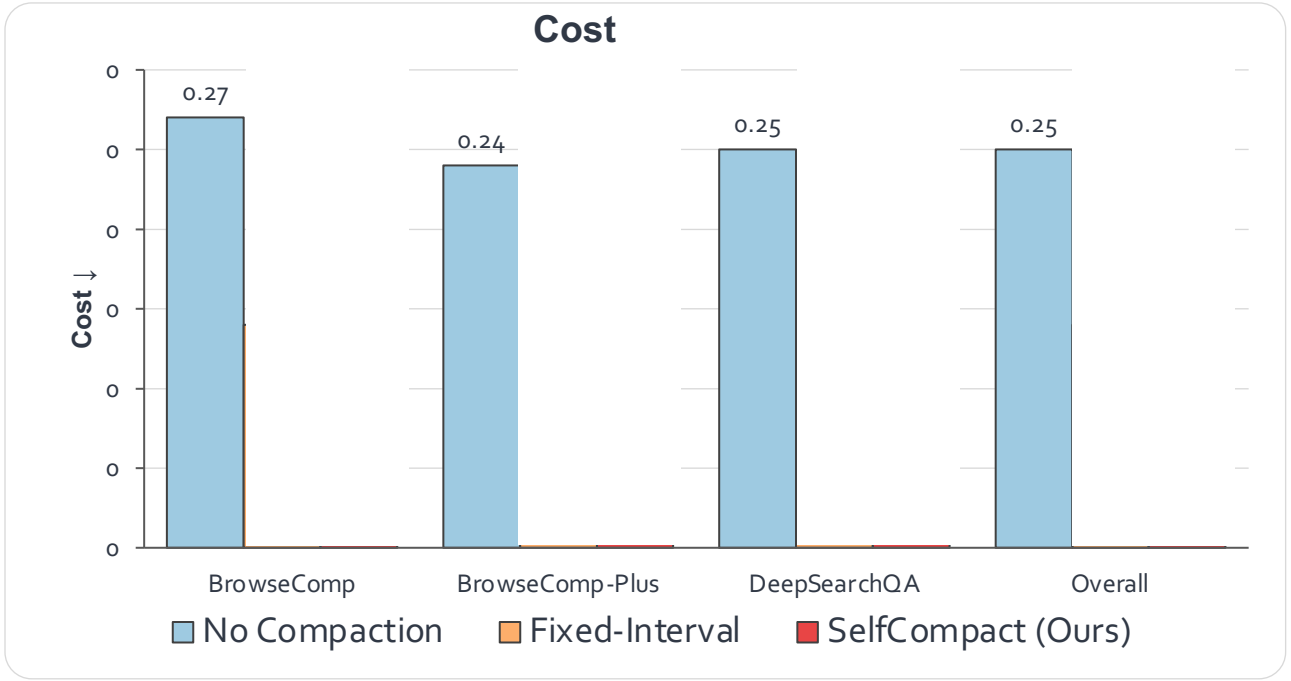


Li et al. Self-Compacting Language Model Agents, 2026.

Empirical Results: Agentic Search

- Mimo-V2-Flash (309B; 15B active)

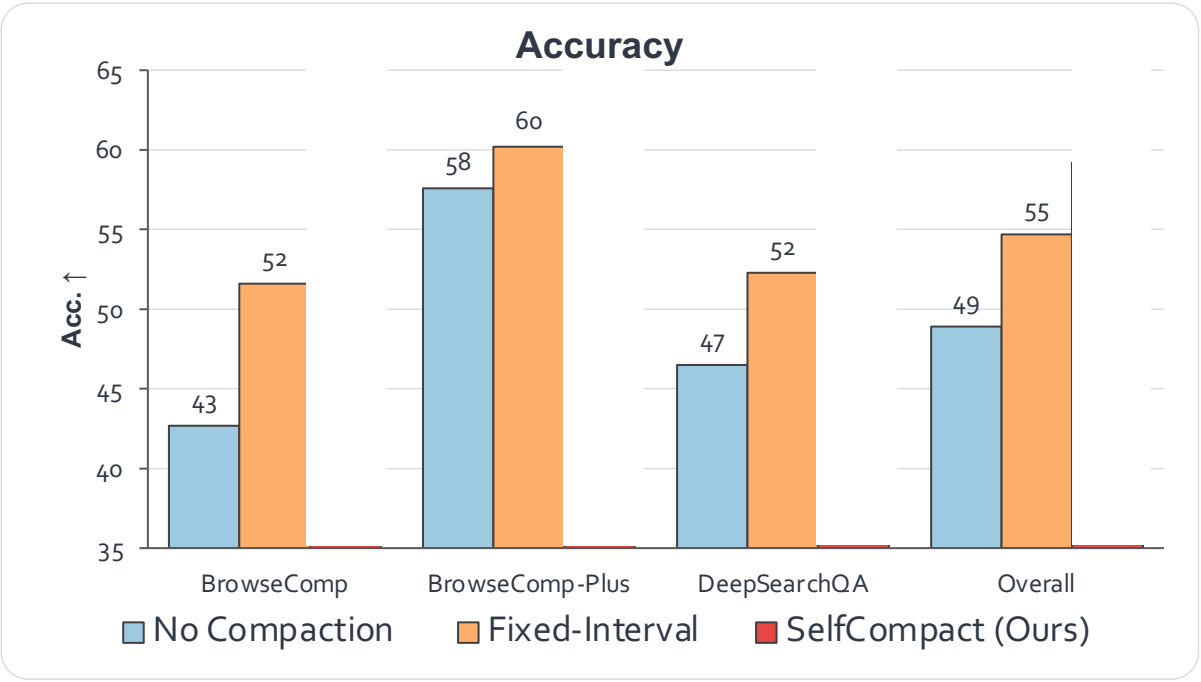


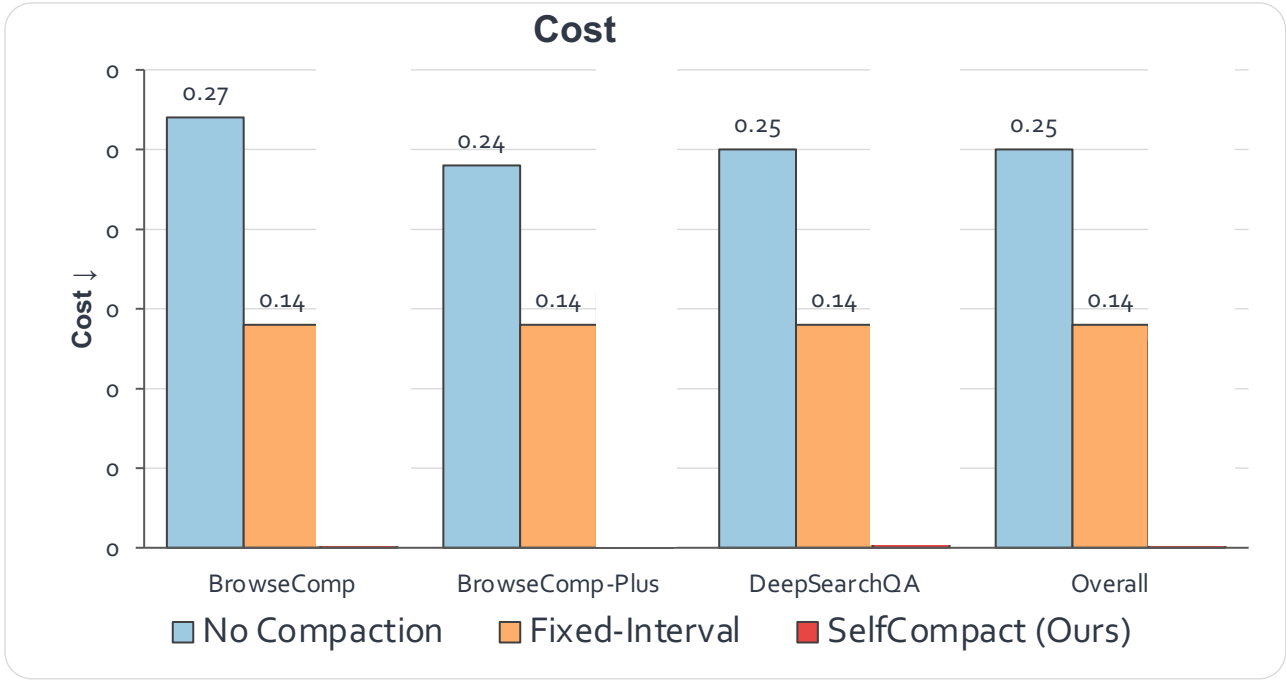


Li et al. Self-Compacting Language Model Agents, 2026.

Empirical Results: Agentic Search

- Mimo-V2-Flash (309B; 15B active)

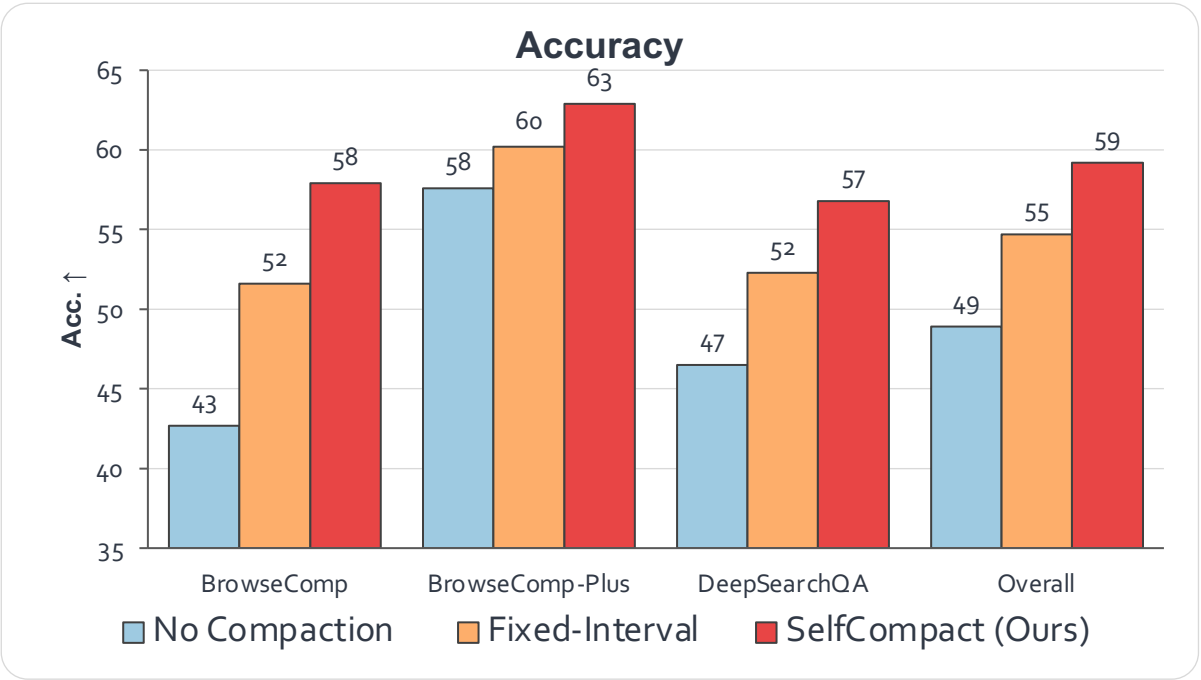


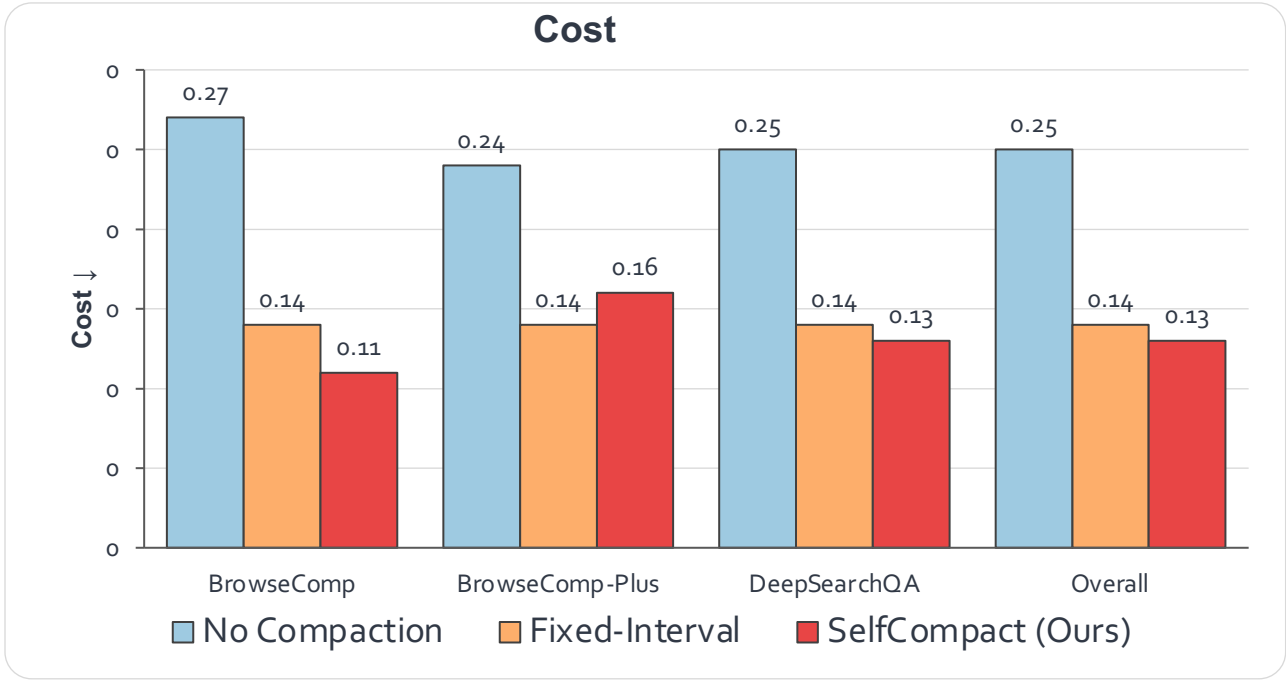


Li et al. Self-Compacting Language Model Agents, 2026.

Empirical Results: Agentic Search

- Mimo-V2-Flash (309B; 15B active)

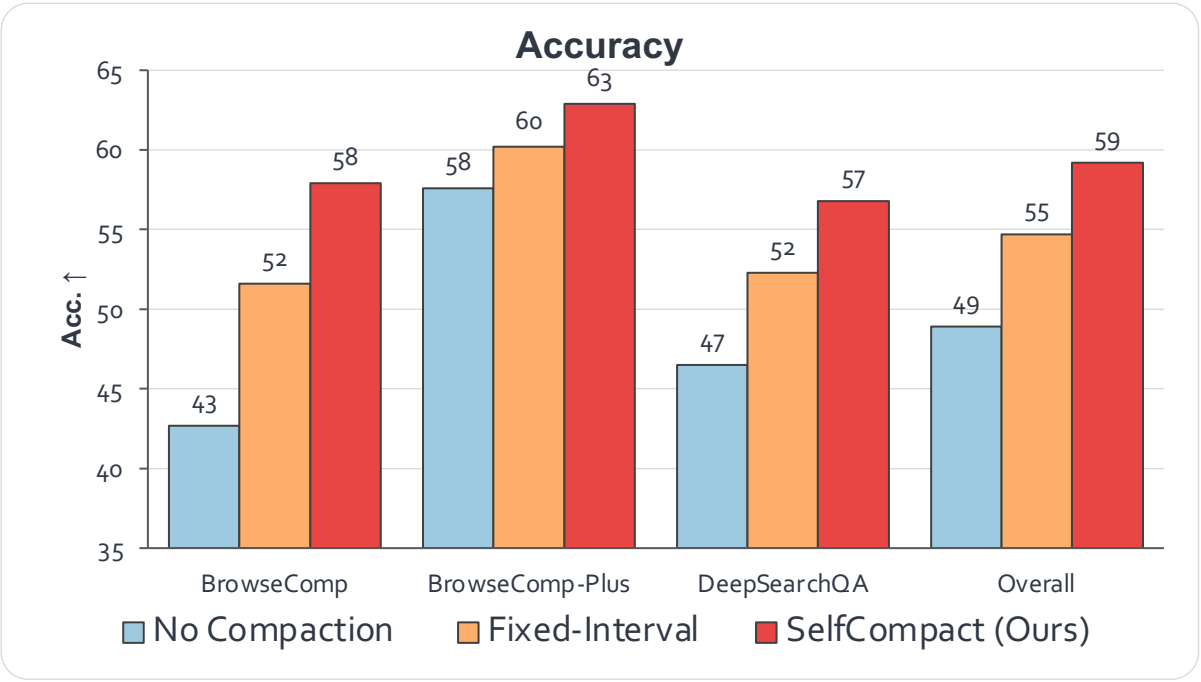


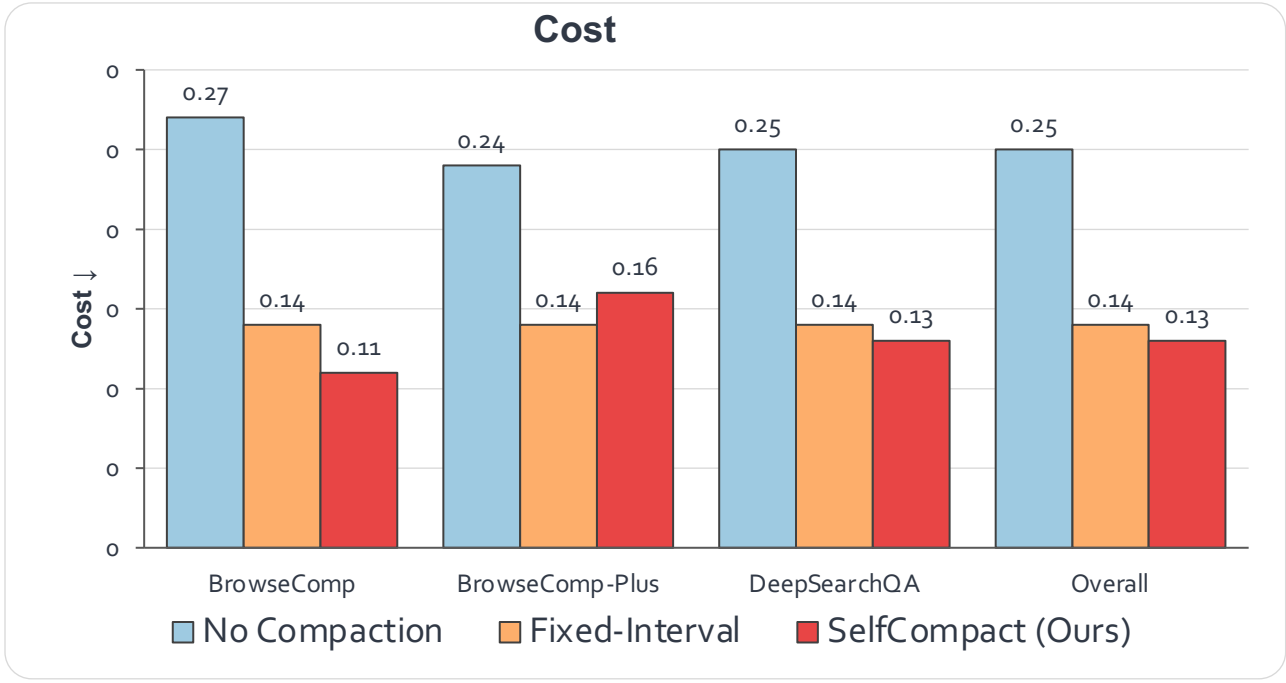


Li et al. Self-Compacting Language Model Agents, 2026.

Empirical Results: Agentic Search

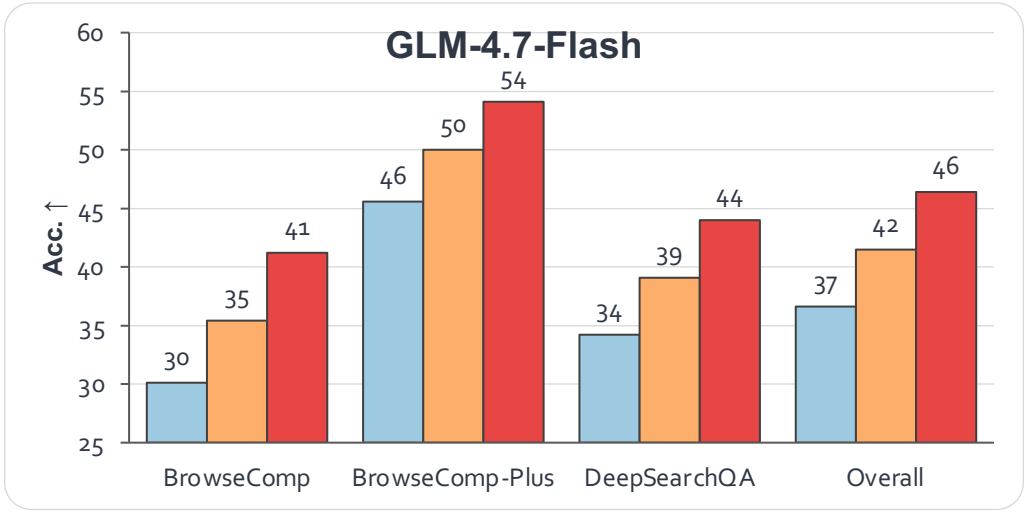
- Mimo-V2-Flash (309B; 15B active)

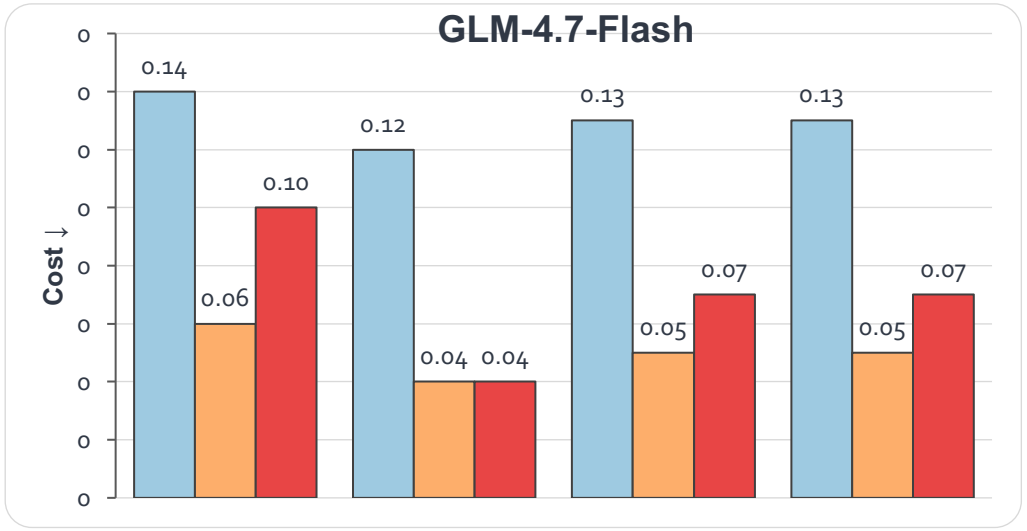


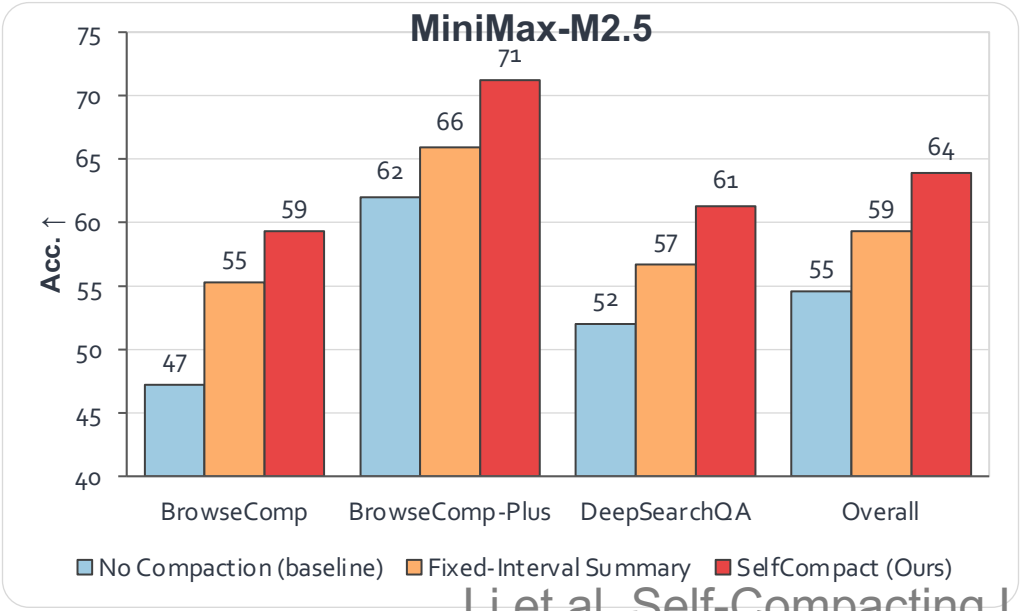


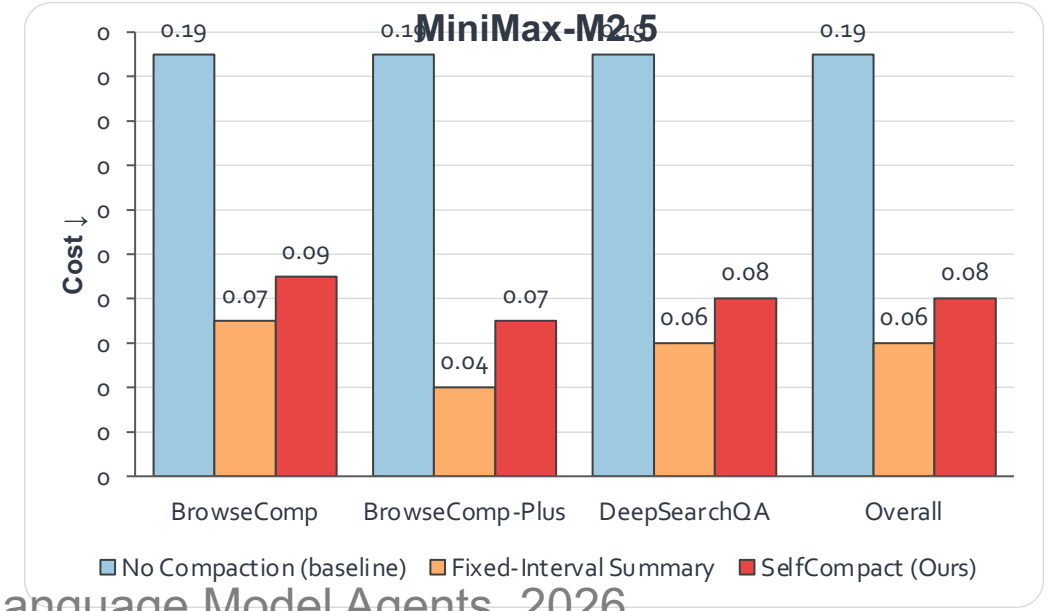
Improves accuracy of fixed-interval compaction while maintaining cost.

Empirical Results: Agentic Search



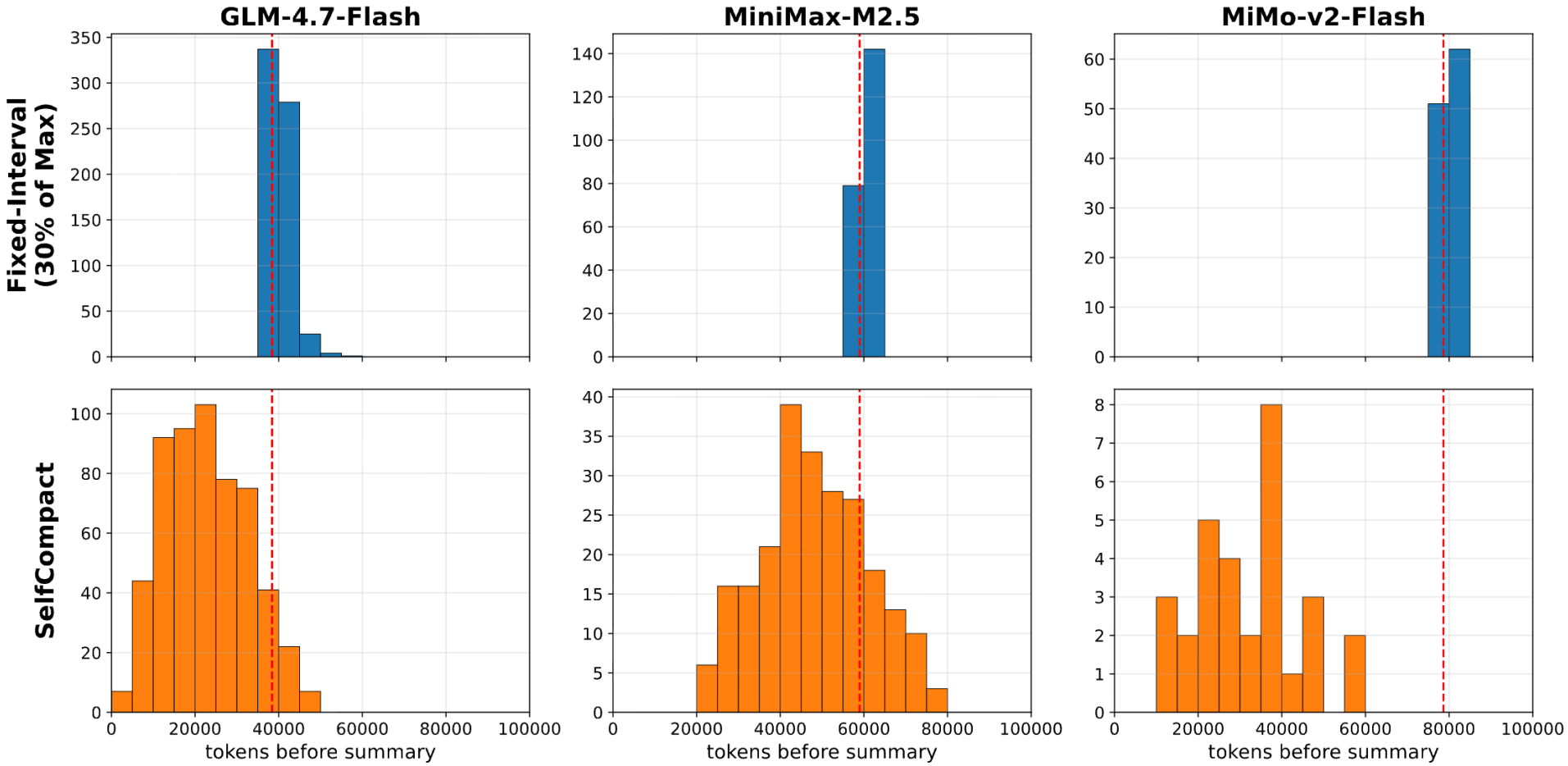






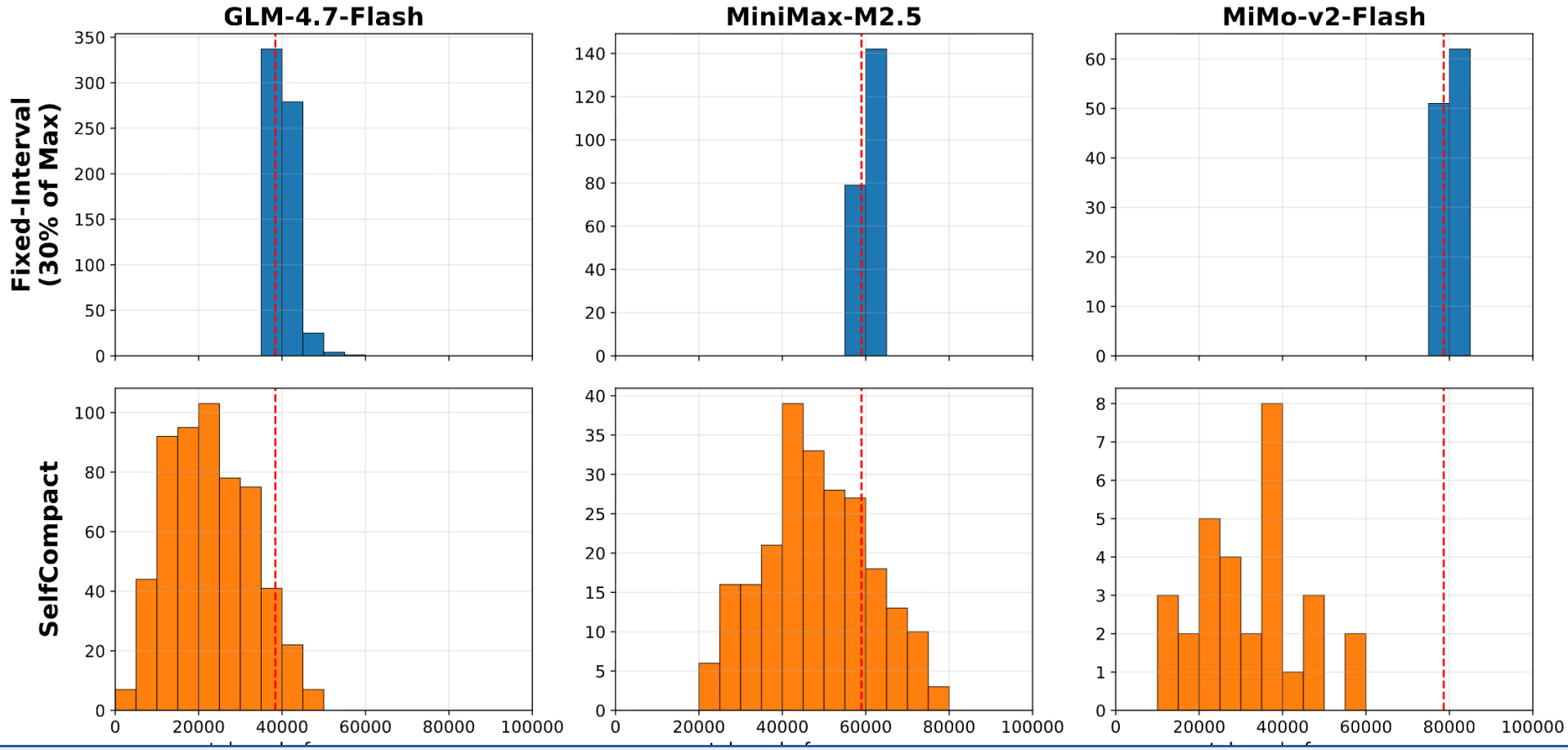
Li et al. Self-Compacting Language Model Agents, 2026.

Why it works: compaction frequency



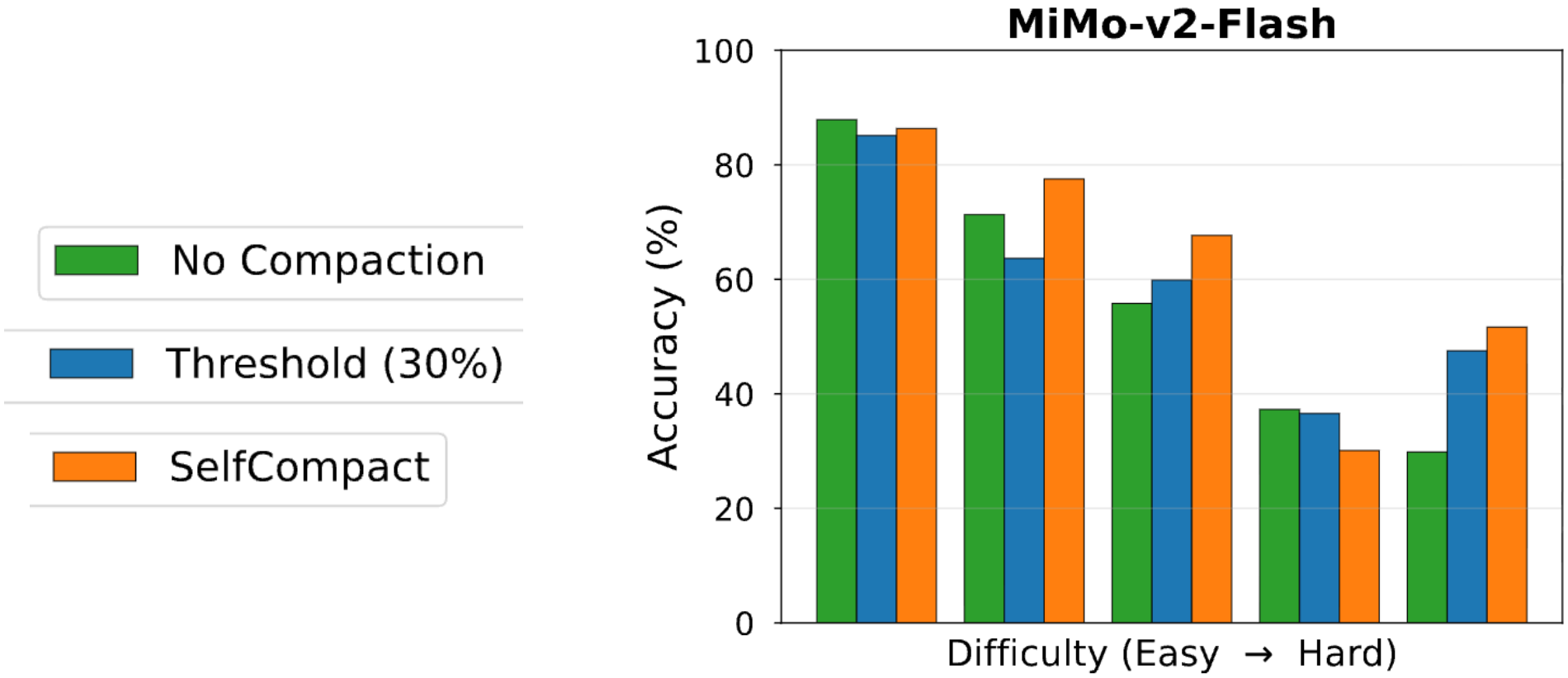
Li et al. Self-Compacting Language Model Agents, 2026.

Why it works: compaction frequency



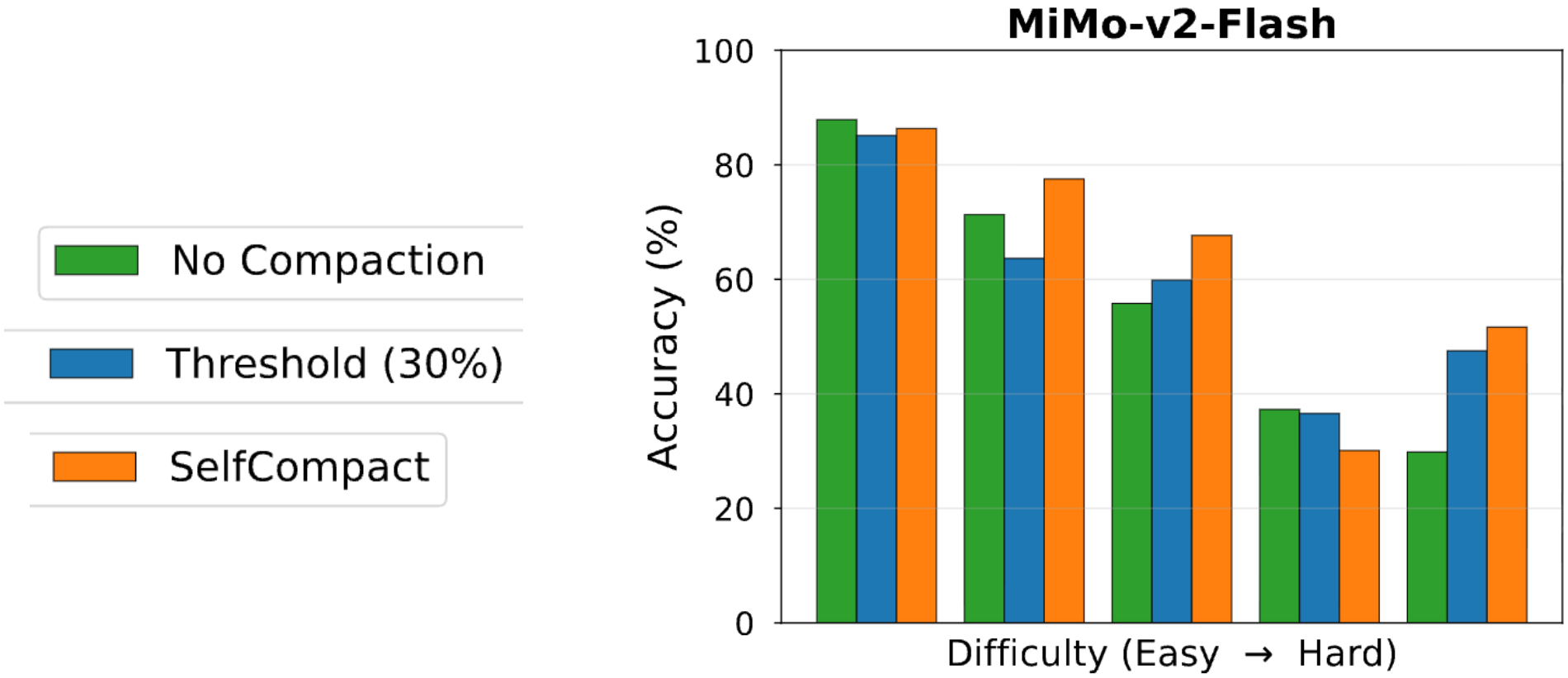
Fixed-Interval triggers at the threshold by construction; SelfCompact spreads them across the budget.

Why it works: ease of problems



Li et al. Self-Compacting Language Model Agents, 2026.

Why it works: ease of problems



SelfCompact matches the others on easy bins and pulls ahead on the hardest bins.

Compacting the content (takeaways)

- Self-Compact: rubric-gated adaptive compaction.
  - Lightweight and training-free.
  - A meta-cognitive capability.
  - Strong empirical results

- Our study is quite simple and likely there is a lot of headroom.
- Call for scientific study of context management.

Compacting the content (takeaways)

- Self-Compact: rubric-gated adaptive compaction.
  - Lightweight and training-free.
  - A meta-cognitive capability.
  - Strong empirical results

- Our study is quite simple and likely there is a lot of headroom.
- Call for scientific study of context management.

Compacting the content (takeaways)

- Self-Compact: rubric-gated adaptive compaction.
  - Lightweight and training-free.
  - A meta-cognitive capability.
  - Strong empirical results

- Our study is quite simple and likely there is a lot of headroom.
- Call for scientific study of context management.

Compacting the content (takeaways)

- Self-Compact: rubric-gated adaptive compaction.
  - Lightweight and training-free.
  - A meta-cognitive capability.
  - Strong empirical results

- Our study is quite simple and likely there is a lot of headroom.
- Call for scientific study of context management.

Thanks to Apple for sponsoring this research!