The H1 test held. The one-quantity bet did not.

1 · TWO FINDINGS, MEASURED APART

MEASURED TOGETHER

Quality loss and rule-breaking come from separate studies.

Rule-breaking 0% to 100%; sums 100% right.

---

2 · THE BET

DID NOT HOLD

One quantity, the operator's share, predicts both.

Share 10.0%: rule held. Share 13.4%: rule broke.

---

3 · THE CONFOUND

SEPARATED

Adding examples changes content and length together.

Neutral text cut to the same token count.

---

4 · PLANNED TEST, H1

HELD

Examples cut the rule's share by more than 10%.

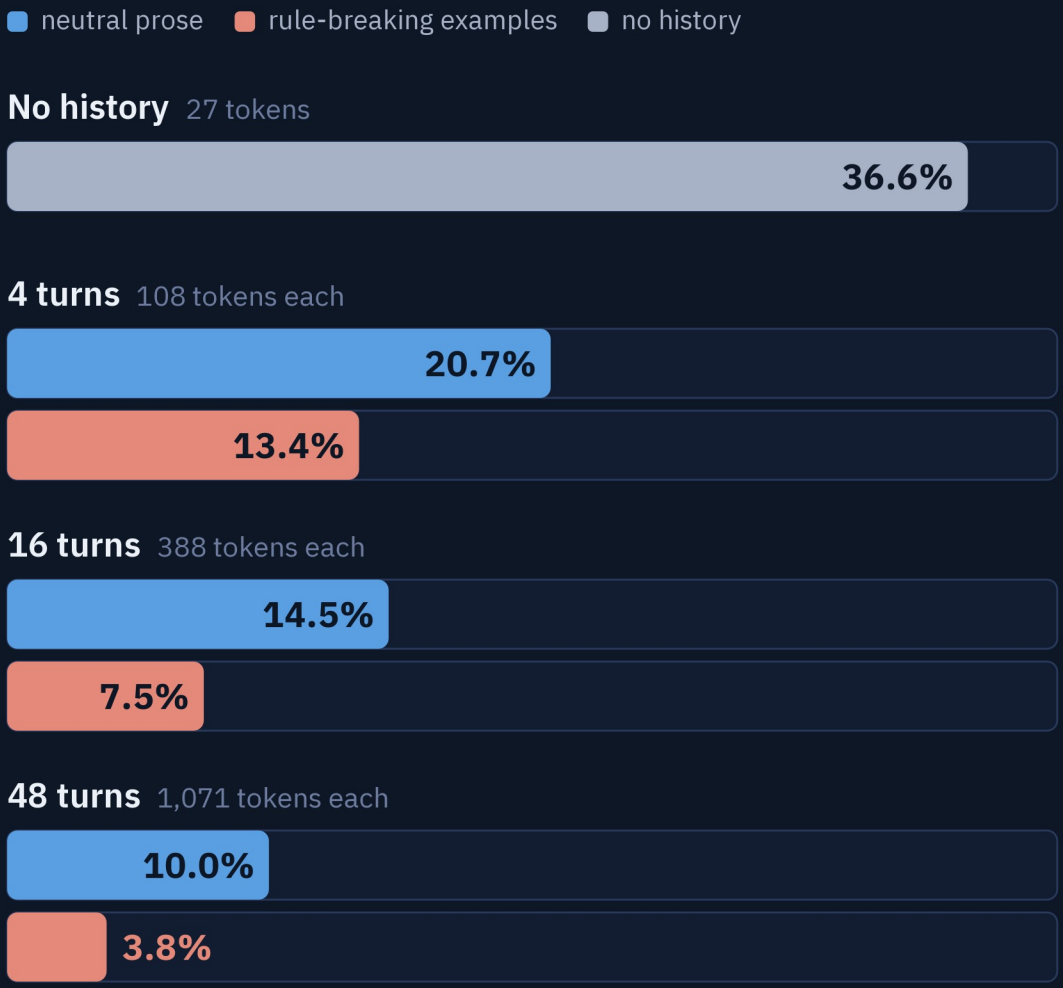
62% lower at 48 turns, 20 of 20 prompts.

Cause check: the rule restated in the last turn took rule-breaking 100% to 0%.

The share fell as predicted; what filled the history decided.

Card: Rock Lambros et al., The Dual-Failure Multiplier (CxBOM). Test: Phi-3.5-mini-instruct, 20 prompts per cell.

H1: rule-breaking examples cut the rule's share 62% at the same length.



Content, not only length, pulls attention off the rule.

Rule's share of attention at the last prompt position, attention sink excluded. Test: Phi-3.5-mini-instruct, 20 prompts per cell.

The bet: a lower share did not mean more rule-breaking.

Neutral prose 1,071 tokens, no examples

RULE'S SHARE OF ATTENTION

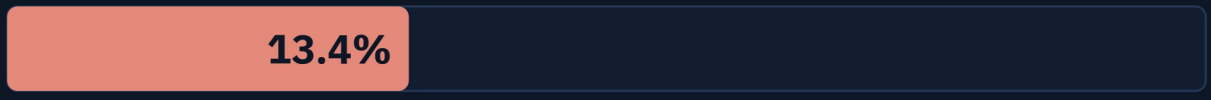


REPLIES BREAKING THE RULE



Four rule-breaking examples 108 tokens

RULE'S SHARE OF ATTENTION



REPLIES BREAKING THE RULE



All eight cells: 0% rule-breaking in every neutral cell, 100% in every example cell, whatever the share.

Rule-breaking followed what filled the history.

Test: Phi-3.5-mini-instruct, 20 prompts per cell.

Cause check: restating the rule took rule-breaking from 100% to 0%.

rule in the system prompt rule restated in the last turn

48 rule-breaking examples 1,071 tokens

RULE'S SHARE OF ATTENTION

3.8%

REPLIES BREAKING THE RULE

100%

Same, plus the rule restated 1,088 tokens

RULE'S SHARE OF ATTENTION

4.3% + 11.9% = 16.2%

REPLIES BREAKING THE RULE

0%

The system prompt's own copy barely moved: 3.8% to 4.3%.

The copy at the point of action did the work.

Test: Phi-3.5-mini-instruct, 20 prompts per cell.

The quality half: 160 of 160 sums were right, so it is still untested.

160 / 160

correct sums, in every cell, at every length

NEXT TESTS

Multi-step math
room to fail

Recall from history
uses the long context

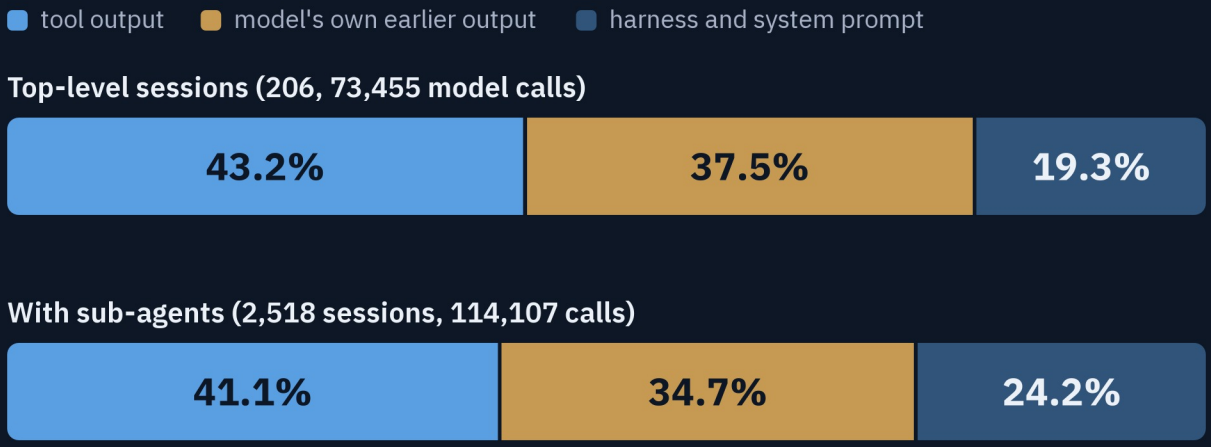
Larger models
other families

The quality half needs a task the model can get wrong as the context fills.

A task that can fail is the next test.

Two-digit additions, 20 per cell, 8 cells. Test: Phi-3.5-mini-instruct, 20 prompts per cell.

The board question: tool output is 41% to 43% of one production context.



System prompt and tool definitions alone: 14.5%. Outside: GitHub Copilot production traces put function-call messages at 28% and the system prompt at 14%.

UPPER BOUND

Tool output includes the operator's own files and command output, not only the open web.

Measure your own: it is all in the transcripts.

One practitioner's Claude Code sessions, 12 May to 29 Oct 2026; user messages excluded. Wolfberg, What Happens in the Model, Table C1.

Read it with the limits: one small model, one harmless rule.

3.8B model
Phi-3.5-mini-instruct

Formatting rule
not a safety rule

20 prompts a cell
greedy decoding

Record arm
uninformative, OR 1.11

Registered before the runs, with every prompt, reply and per-layer share published.

Worth repeating at scale, with a safety rule.

Test: Phi-3.5-mini-instruct, 20 prompts per cell.

Everything is on the page: paper, data and the run script.

Paper
HTML and PDF

Registration
written before the runs

Raw rows
every prompt and reply

Run script
Phi-3.5-mini, any GPU

wolfberg.ai/papers/
rule-share-is-not-the-lever

Hypothesis and H1 design: Rock Lambros et al., the Dual-Failure Multiplier (CxBOM).

Wolfberg research note, 29 October 2026