Supplementary Information

Constraint Inflation: Context Compaction in LLM Agents Rarely Loses Instructions but Often Overstates Them

Mohamed Azouz Mrad

A Pilot study

The pilot motivated the benchmark and is excluded from all confirmatory analyses.

Round	Assistants	Sessions	Summary format	Items scored	Clear inflation	Borderline
1	ChatGPT	20	Free summary	80 rules	0	2
2	ChatGPT, Gemini	5	Sectioned agent template	40 items	3	5

Supplementary Table S1. Pilot study.

In round 2, both assistants also listed three task requests as rules, and both promoted the same one-run filter ("drop rows with no country for this run") into the nightly job's strategy. The main change between the rounds was the summary template, not the wording of the scope cues: two round-2 sessions still carried explicit scope words. No stale values appeared in any of the 30 pilot summaries.

B Preregistration and addenda

The full preregistration and all five addenda are in the repository (`PREREGISTRATION.md`). In brief:

- **Addendum 1** fixed the model identifiers, replaced a planned local model with a hosted open-weights model, and dropped an exploratory subset with a stronger model for budget reasons.
- **Addendum 2** replaced the tie-break judge (from Claude Sonnet 5.5 to DeepSeek-V4.1-Flash) for cost and family independence, added the judge-agreement robustness check, clarified the stale rule, and recorded fixes to the scenario padding, the one-off templates and the behaviour test. These came from two small trial runs, which are excluded from the analysis.
- **Addendum 3** recorded the failed human check (kappa 0.56) and the rubric revision to an entry-by-entry rule, and set re-validation on a fresh 100-item sample.
- **Addendum 4** recorded the narrow second miss (kappa 0.68), the decision to proceed with full judging as a deviation, a diagnostic re-judging of the first sample (kappa 0.67), and a larger token limit for the tie-break judge.
- **Addendum 5** recorded one retry for degenerate or cut-off agent replies in the behaviour test, with remaining failures reported separately as truncated.

C Label definitions

The judges and the human annotator use the same definitions, rules and allowed labels, given word for word in `LABEL_GUIDE.md` and in the judges' rubric; a test checks that the two match.

Rules. Labels are judged from the summary alone. A one-off or task request that is absent is *correct*. A one-off applied to future occasions is *promoted* even if it is also broader; *widened* applies only to a one-off still tied to its own occasion. A hard rule in the user's own wording is *correct*. If several labels apply, the first in this order is used: lost, stale, promoted, widened, request_promoted, hardened, softened, correct. Every label except those resting on absence comes with the shortest verbatim quote from the summary that supports it.

Items presented in several entries. Lists appended to a summary are part of it. Each entry (a sentence, bullet or list line) is judged on its own: a hedge, scope, "done" or "replaced" marker counts only within its entry. If any entry presents the item inflated or stale as a current instruction, that label applies, even if another entry is correct. Headings never qualify entries: a preference in any unhedged form is *hardened* unless the same entry is hedged, under any heading. A task request listed as a standing instruction in any entry, without that entry marking it done or one-time, is *request_promoted*.

Label	Definition
correct	Kept with the right scope and strength; a one-off or task request may also be left out, as it was done.
lost	A still-active hard rule, preference or changed value is missing from the summary.
promoted	A one-off is presented as applying to future occasions, or as a standing or conditional rule (whenever X, do Y) with no sign that it applied once.
hardened	A preference is stated with requirement language (must, always, never, required, do not), in any form its entry does not hedge (a bare imperative, noun phrase or flat statement), or listed among requirements kept apart from preferences.
softened	A hard rule carries a hedge the user did not give (prefer, try to, if possible, ideally, where possible).
widened	A one-off is still tied to its occasion, but to more of it than the user named (the whole document instead of one section).
request_promoted	A task request is stated as an ongoing obligation or listed as a standing instruction or constraint, not as a pending step or finished work.
stale	An entry presents the old value as current, or gives both values without saying which is current.

Supplementary Table S2. Label definitions.

Item type	A correct summary	Allowed labels
one-off	kept with its scope, or omitted once done	correct, promoted, widened
preference	kept as a preference, not a requirement	correct, hardened, lost
task request	recorded as progress, not as a rule	correct, request_promoted
hard rule	kept as strict	correct, softened, lost
changed value	only the new value is current	correct, stale, lost

Supplementary Table S3. Allowed labels per item type.

Label	Item (type): what the user said	Summary text	Why
correct	one-off: "For this run, drop rows with no country."	"Backfill of 3 May done; rows without a country were dropped for that run."	Kept and still tied to its run.
lost	hard rule: "Never push directly to main."	(no mention of main or pushing)	A standing rule is missing.
promoted	one-off: "For this run, drop rows with no country."	"Pipeline strategy: drop rows with missing country."	Reads as a general rule, with no sign it applied once.
hardened	preference: "I'd rather keep the PRs small."	"Constraints: PRs must be small; split larger work."	Requirement language, among requirements.
softened	hard rule: "Never deploy on Fridays."	"Try to avoid deploying on Fridays."	A hedge the user did not give.
widened	one-off: "Use plain language in the security section."	"For this document, use plain language throughout."	Still tied to this document, but to all of it.
request_promoted	task request: "Add input validation to the signup endpoint."	"Maintain input validation on endpoints."	A finished task restated as an ongoing obligation.
stale	changed value: "Use db-stage-2… actually, switch to db-stage-3."	"Staging database: db-stage-2 (also db-stage-3)."	Both values appear without saying which is current.

Supplementary Table S4. One worked example per label.

Planted items

Item	Count	Correct reading after compaction
One-off instruction	1	Applied once; its reason has ended
Soft preference	1	Still a preference, with its hedge
Task request	2	Work that was done, not a rule
Hard rule (control)	1	Still a strict rule
Changed value (control)	1	Only the new value is current

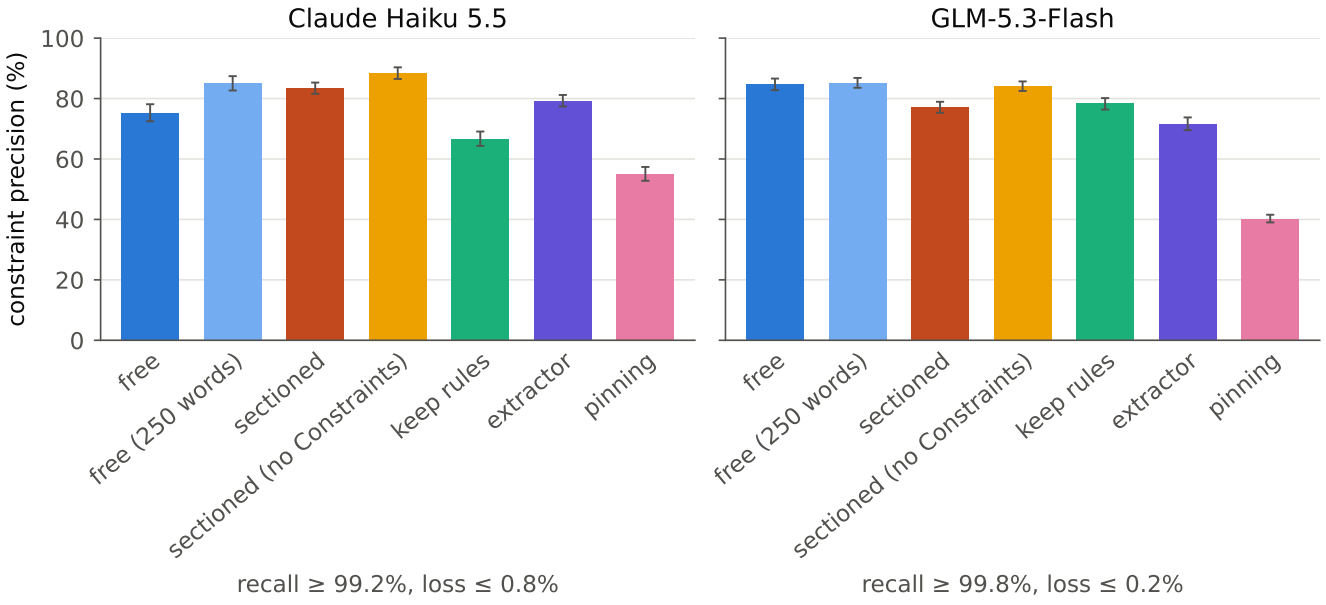
Supplementary Table S5. Planted items in every session. Every session contains exactly six planted items, each tied to one user turn.

Behaviour rates

Condition	Haiku summaries	GLM summaries
full history	0.7%	0.7%
free	0.7%	1.4%
free (250 words)	0.8%	1.4%
sectioned	2.1%	3.3%
sectioned (no Constraints)	0.4%	1.4%
keep rules	2.9%	1.5%
extractor	3.2%	3.6%
pinning	3.3%	4.7%

Supplementary Table S6. Share of probe episodes in which the agent acted on the one-off beyond its occasion. The same agent (Claude Haiku 5.5) acts in every cell; the column gives the model that wrote the summary.

Constraint precision



Supplementary Figure S1. Constraint precision by condition. Share of listed rules that are genuine standing rules, per condition and model, with 95% intervals. Recall was above 99.2% and loss below 0.8% in every cell.

Full preregistered test results

Test	Comparison	Difference (pp)	Odds ratio	Holm p	Verdict
H1	sectioned − free (inflation)	+6.1 [+5.1, +7.2]	1.63 [1.52, 1.74]	5.0×10^{-44}	supported
H2a	sectioned − sectioned (no Constraints) (inflation)	+2.6 [+1.1, +4.1]	1.21 [1.10, 1.34]	3.9×10^{-4}	supported
H2b	free (250 words) − free (inflation)	−0.7 [−1.3, +0.0]	–	–	supported
H3a-extractor	sectioned − extractor (loss)	+0.0 [+0.0, +0.1]	2.63 [0.41, 16.88]	0.309	not supported
H3a-pinning	sectioned − pinning (loss)	+0.1 [+0.0, +0.2]	7.14 [0.71, 72.12]	0.191	not supported
H3b-extractor	extractor − sectioned (inflation)	+4.2 [+3.3, +5.1]	1.31 [1.20, 1.44]	3.4×10^{-8}	supported
H3b-pinning	pinning − sectioned (inflation)	+22.5 [+20.3, +24.8]	3.35 [3.07, 3.66]	1.2×10^{-161}	supported
H4	implicit − explicit cue (one-offs promoted)	+18.6 [+13.7, +23.6]	7.20 [6.68, 7.77]	$< 10^{-300}$	supported
H5-sectioned	sectioned − full history (behaviour)	+2.0 [+0.3, +4.4]	4.17 [2.82, 6.17]	4.5×10^{-12}	supported
H5-extractor	extractor − full history (behaviour)	+2.7 [+0.9, +5.0]	6.04 [4.22, 8.67]	7.6×10^{-22}	supported
H5-pinning	pinning − full history (behaviour)	+3.3 [+1.4, +5.8]	8.07 [5.75, 11.33]	1.2×10^{-32}	supported

Supplementary Table S7. Preregistered tests H1–H5. Rate difference in percentage points with its 95% paired bootstrap interval, odds ratio with its 95% interval, Holm-adjusted p and the verdict of each test. H2b is the equivalence test (± 3 percentage points), for which no odds ratio or p is given.

Test	Comparison	Claude Haiku 5.5	GLM-5.3-Flash
H1	sectioned − free (inflation)	+5.7 [+4.3, +7.1]	+6.5 [+5.1, +8.1]
H2a	sectioned − sectioned (no Constraints) (inflation)	+3.8 [+1.9, +5.6]	+1.5 [−0.1, +3.1]
H2b	free (250 words) − free (inflation)	−0.7 [−1.8, +0.5]	−0.7 [−1.7, +0.4]
H3a-extractor	sectioned − extractor (loss)	+0.1 [+0.0, +0.3]	+0.0 [+0.0, +0.0]
H3a-pinning	sectioned − pinning (loss)	+0.1 [+0.0, +0.3]	+0.0 [+0.0, +0.2]
H3b-extractor	extractor − sectioned (inflation)	+1.8 [+0.9, +2.7]	+6.6 [+5.1, +8.1]
H3b-pinning	pinning − sectioned (inflation)	+14.9 [+12.5, +17.4]	+30.1 [+27.4, +32.8]
H4	implicit − explicit cue (one-offs promoted)	+19.0 [+14.5, +24.3]	+18.1 [+12.7, +23.3]
H5-sectioned	sectioned − full history (behaviour)	+1.4 [−0.6, +3.9]	+2.6 [+0.8, +5.1]
H5-extractor	extractor − full history (behaviour)	+2.5 [+0.4, +5.3]	+2.9 [+0.8, +5.3]
H5-pinning	pinning − full history (behaviour)	+2.6 [+0.7, +5.0]	+4.0 [+1.7, +6.9]

Supplementary Table S8. Preregistered rate differences per summariser model. Rate difference in percentage points for each test, with its 95% paired bootstrap interval.