Memory error card

Controlled speaker filter · gpt-oss-120b · random links at d = 0.8
200 workplace questions / 5 histories; 80 social questions / 41 histories.
Three draws and four option orders. Message text and names stay fixed.
Comparator: correct index. Budget: 2,048 tokens; shortlist: 64 messages.

Measure	Workplace	Social
Exposure	95.50%	84.08%
Context change	81.00%	73.37%
Evidence displacement	22.83%	3.25%
Churn	19.25% churn. Gains 6.04%, losses 7.79%. Net -1.75 points.	4.07% churn. Gains 1.05%, losses 2.57%. Net -1.52 points.
Conditional change	Loss among right: 14.43%. Gain among wrong: 13.13%.	Loss among right: 3.20%. Gain among wrong: 5.38%.
Index headroom	-0.62 points	+0.51 points
No-memory agreement	63.90% of stable valid answers after context change. Stratum mass 61.75%.	29.74% of stable valid answers after context change. Stratum mass 69.31%.
Repeated requests	3/24 selected repeated requests disagree.	5/24 selected repeated requests disagree.
Loss concentration	75/200 questions have a loss. Top 20: 47.59% of loss; top question: 4.28%.	5/80 questions have a loss. Top 8: 100.00% of loss; top question: 78.95%.

Rates retain equal history weights and original mass before conditioning.
Churn includes service variation and one invalid workplace answer.
Unchanged contexts reuse answers. Net = gain minus loss.
Headroom does not bound harm. No-memory agreement does not prove non-use.
Replay counts use selected requests across conditions, not this full population.
Synthetic histories; imperfect evidence labels; descriptive saved-data analyses.
Point estimates only. The paper gives social history-bootstrap intervals and workplace omit-one-history sensitivity ranges.
Only 5 social questions lose answers, so the top-8 share must be 100%.
Concentration uses weighted losses; its independent simulation omits dependence.

Memory error card template

System and version: _____ Date: _____
Intervention: _____ Comparator: _____
Workload and scope: _____
Questions / histories / draws / orders: _____

Measure	Population or measure	Result / uncertainty
Exposure	Searches that reach an additional person	
Context change	Pairs with different serialized contexts	
Evidence displacement	Labeled messages gained or lost	
Churn	Gains, losses, wrong switches, invalid changes	
Conditional change	Loss among right answers; gain among wrong answers	
Index headroom	Correct-index minus unfiltered accuracy	
No-memory agreement	Stable answers matching no-memory choice	
Repeated requests	Changed answers under identical input	
Loss concentration	Questions with loss; share in the top 10% and top question	

Record pairing keys, missing cells, invalid responses, and option mapping.
State the weighting rule and the unit used for uncertainty.
Give distinct questions and histories separately from repeated comparisons.
State repeat selection and all changes to the prompt or serving endpoint.
Record source IDs, code versions, and the analysis receipt.
For concentration references, state eligibility, weighting, and dependence.