PARTIAL QUANTUM SEARCH · LLM EVALUATION

Failure Analysis Across Models and Questions

Where each model breaks, why, and how often — Stage 1 (one-shot) and Stage 2 (robustness under challenge).

Ustunpasa Igdir · June 28, 2026

Data provenance

This analysis is computed directly from the project database (*stage1.db*). **Stage 1** covers the current reviewed one-shot set: **82 runs · 8 models · 4 scenarios** (Gemini and Qwen3 8B already excluded as unreliable). **Stage 2** covers **189 same-session follow-ups · 7 models · 3 scenarios**; DeepSeek Reasoner appears in Stage 1 as a tested model but in Stage 2 serves as the independent second-opinion grader, so it is not scored as a Stage 2 contestant.

Note: the published Stage 1 report card uses an earlier 120-run snapshot (which still included Gemini and Qwen3 8B). The per-model / per-question breakdowns below come from the current 82-run database — the only set with granular human-reviewed labels — so totals differ from that headline by design.

Failure definitions

Stage 1 failure = a reviewed run with final verdict CRITICAL_ERROR or NUMERIC_ERROR (UNUSABLE/incomplete runs are excluded from failure counts). **Stage 2 failure** = a follow-up with human verdict FAIL (PARTIAL is tracked separately as a softer signal).

1 · Headline findings

Robustness tracks capability tier, not just one-shot accuracy. The three strongest models — Perplexity Sonar Reasoning Pro (26/27), Perplexity Sonar (24/27) and Mistral Medium 3.5 (19/27) — recorded **zero** Stage 2 FAILs. All Stage 2 hard failures are concentrated in the small 8–12B group plus, notably, the larger Llama 3.3 70B on one scenario.

Llama 3.1 8B is the most fragile model overall. It passed only 4/27 Stage 2 follow-ups and failed in all three scenarios; its worst was BASE (9 FAILs), most often by *introducing a new error* when challenged rather than repeating the old one.

Each scenario has a signature failure reason. SCALE failures are dominated by C_PARALLEL_AS_SEQUENTIAL (treating parallel local search as sequential); CONSISTENCY failures by C_NO_VALID_REPAIR_PROPOSED (no valid repair for the impossible equal-block setup); BASE failures are mixed (parallel-as-sequential, invalid-formula rejection, and new errors under pressure).

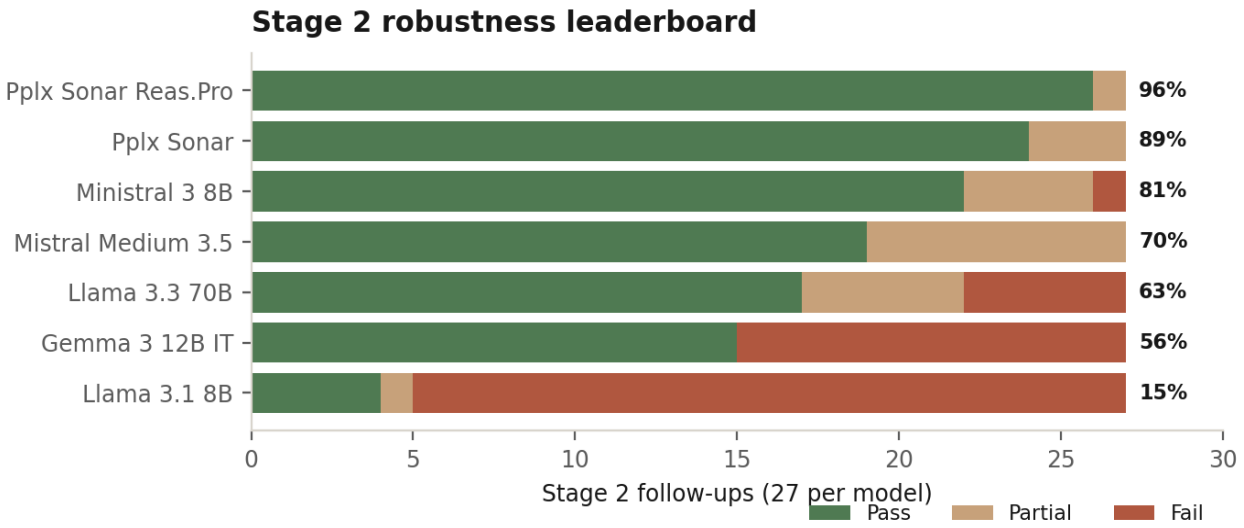
One-shot success does not guarantee robustness. Llama 3.3 70B passed every Stage 1 question yet failed BASE robustness 5/9 in Stage 2. Gemma 3 12B passed CONSISTENCY one-shot but failed it 3 times under challenge. Correct first answers and stable reasoning are different abilities.

The Stage 2 grader had itself failed in Stage 1. DeepSeek Reasoner — used as the independent second-opinion grader in Stage 2 — was the model that failed CONSISTENCY 3/3 in Stage 1 (no valid repair proposed). It grades follow-ups for scenarios where its own one-shot reasoning was weak; human review remains the final authority.

A correct first answer predicts robustness far better than a recoverable one. Answers that passed Stage 1 held up 75% of the time under challenge (114/153). Answers that had failed — whether *conceptually* (38% recovered, 9/24) or *numerically* (33%, 4/12) — recovered at roughly half that rate when given a targeted correction. Conceptual and numeric failers recovered about equally poorly.

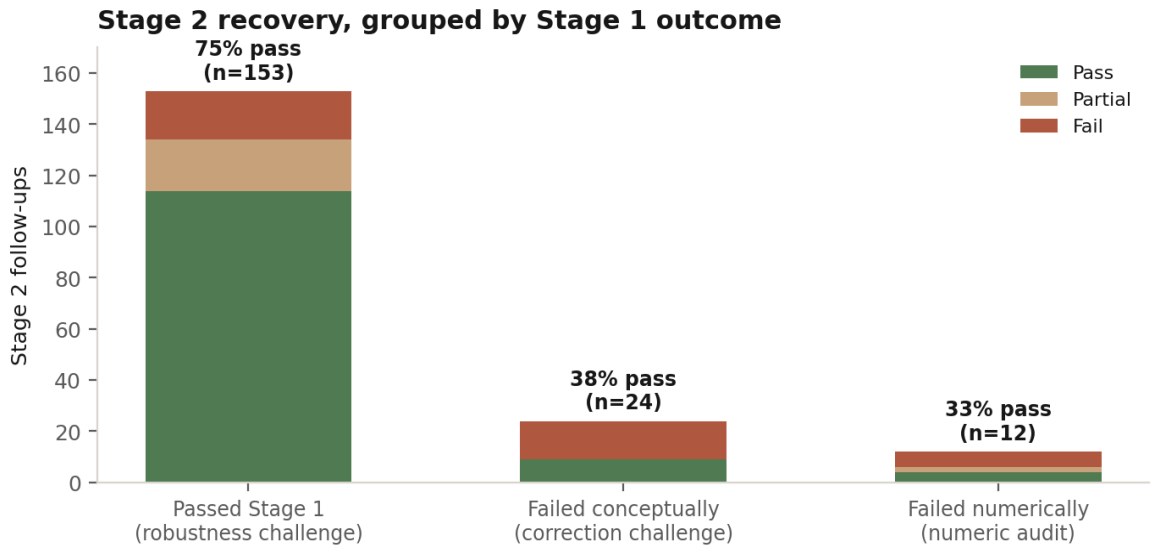
2 · Visual summary

2.1 · Robustness leaderboard



2.2 · Stage 2 recovery, grouped by Stage 1 outcome

The comparison you asked for: how each follow-up fared in Stage 2, split by what its Stage 1 parent answer had been — passed, failed conceptually (CRITICAL_ERROR), or failed numerically (NUMERIC_ERROR).



Stage 1 outcome	Stage 2 challenge	Pass	Partial	Fail	n	Pass %
Passed	robustness	114	20	19	153	75%
Failed conceptually	conceptual correction	9	0	15	24	38%
Failed numerically	numeric audit	4	2	6	12	33%

Read with care: groups got *different* challenge types (defend vs. repair), so "pass" isn't identical across them — the signal is directional. Failed-group samples are small (n=24, n=12): evidence for review, not estimates.

3 · Stage 1 — where each model failed most

For each model: the scenario it failed most, how many times, and the reason(s) recorded for those failures. Models with no CRITICAL/NUMERIC failures are listed as clean.

Model	Class	Worst question	Fails	Top reason(s) for that question
DeepSeek Reasoner	larger / cloud	CONSISTENCY	3	C_NO_VALID_REPAIR_PROPOSED ×2, (unlabeled) ×1
Gemma 3 12B IT	small 8–12B	SCALE	3	C_PARALLEL_AS_SEQUENTIAL ×3
Llama 3.1 8B Instruct	small 8–12B	BASE	3	INVALID_FORMULA_REJECTION ×2, N_MISSING_NUMERIC_RESULT ×1, N_TOTAL_VALUE_ERROR ×1
Llama 3.3 70B Instruct	larger / cloud	—	0	no CRITICAL/NUMERIC failures
Ministral 3 8B	small 8–12B	BASE	3	C_PARALLEL_AS_SEQUENTIAL ×3
Mistral Medium 3.5	larger / cloud	—	0	no CRITICAL/NUMERIC failures
Perplexity Sonar	larger / cloud	SCALE	1	N_TOTAL_VALUE_ERROR ×1
Perplexity Sonar Reasoning Pro	larger / cloud	—	0	no CRITICAL/NUMERIC failures

4 · Stage 1 — per question, models ranked most → least failed

BASE — baseline parallel search

#	Model	Class	Fails	Reasons
1	Llama 3.1 8B Instruct	small 8–12B	3	INVALID_FORMULA_REJECTION ×2, N_MISSING_NUMERIC_RESULT ×1, N_TOTAL_VALUE_ERROR ×1
2	Ministral 3 8B	small 8–12B	3	C_PARALLEL_AS_SEQUENTIAL ×3

CONSISTENCY — inconsistent equal blocks

#	Model	Class	Fails	Reasons
1	DeepSeek Reasoner	larger / cloud	3	C_NO_VALID_REPAIR_PROPOSED ×2, (unlabeled) ×1

SCALE — astronomical-scale correction term

#	Model	Class	Fails	Reasons
1	Gemma 3 12B IT	small 8–12B	3	C_PARALLEL_AS_SEQUENTIAL ×3
2	Llama 3.1 8B Instruct	small 8–12B	2	NUMERIC_ERROR ×2
3	Perplexity Sonar	larger / cloud	1	N_TOTAL_VALUE_ERROR ×1

COUNT — counting variant

No CRITICAL/NUMERIC failures recorded for this question.

5 · Stage 2 — where each model failed most (under challenge)

Stage 2 challenges each Stage 1 answer 3× in the same session. Below: each model's worst scenario by FAIL count, the reason(s), and its overall follow-up record. DeepSeek is excluded (it is the grader).

Model	Class	Worst Q	FAILs	Top reason(s)	Overall (P/Pt/F of 27)
Gemma 3 12B IT	small 8–12B	SCALE	9	C_PARALLEL_AS_SEQUENTIAL ×9	15/0/12
Llama 3.1 8B Instruct	small 8–12B	BASE	9	NEW_ERROR ×6, (blank) ×2, F_PARTIAL_CORRECTION_ONLY ×1	4/1/22
Llama 3.3 70B Instruct	larger / cloud	BASE	5	(blank) ×2, COUNTERFACTUAL_MISSING ×1, COUNTERFACTUAL_OFF ×1, F_PARTIAL_CORRECTION_ONLY ×1	17/5/5
Ministral 3 8B	small 8–12B	CONSISTENCY	1	C_NO_VALID_REPAIR_PROPOSED ×1	22/4/1
Mistral Medium 3.5	larger / cloud	—	0	no hard FAILs	19/8/0
Perplexity Sonar	larger / cloud	—	0	no hard FAILs	24/3/0
Perplexity Sonar Reasoning Pro	larger / cloud	—	0	no hard FAILs	26/1/0

6 · Stage 2 — per question, models ranked most → least failed

BASE — baseline parallel search

#	Model	Class	FAILs	Reasons
1	Llama 3.1 8B Instruct	small 8–12B	9	NEW_ERROR ×6, (blank) ×2, F_PARTIAL_CORRECTION_ONLY ×1
2	Llama 3.3 70B Instruct	larger / cloud	5	(blank) ×2, COUNTERFACTUAL_MISSING ×1, COUNTERFACTUAL_OFF ×1, F_PARTIAL_CORRECTION_ONLY ×1

CONSISTENCY — inconsistent equal blocks

#	Model	Class	FAILs	Reasons
1	Llama 3.1 8B Instruct	small 8–12B	7	C_NO_VALID_REPAIR_PROPOSED ×7
2	Gemma 3 12B IT	small 8–12B	3	C_NO_VALID_REPAIR_PROPOSED ×3
3	Ministral 3 8B	small 8–12B	1	C_NO_VALID_REPAIR_PROPOSED ×1

SCALE — astronomical-scale correction term

#	Model	Class	FAILs	Reasons
1	Gemma 3 12B IT	small 8–12B	9	C_PARALLEL_AS_SEQUENTIAL ×9
2	Llama 3.1 8B Instruct	small 8–12B	6	F_COUNTERFACTUAL_UPDATE_FAIL ×3, F_NUMERIC_AUDIT_FAIL ×3

7 · Non-trivial comparisons

7.1 · Robustness leaderboard (Stage 2 pass rate)

#	Model	Class	Pass	Partial	Fail	Pass %
1	Perplexity Sonar Reasoning Pro	larger / cloud	26	1	0	96%
2	Perplexity Sonar	larger / cloud	24	3	0	89%
3	Ministral 3 8B	small 8–12B	22	4	1	81%
4	Mistral Medium 3.5	larger / cloud	19	8	0	70%
5	Llama 3.3 70B Instruct	larger / cloud	17	5	5	63%
6	Gemma 3 12B IT	small 8–12B	15	0	12	56%
7	Llama 3.1 8B Instruct	small 8–12B	4	1	22	15%

7.2 · One-shot vs. robustness — same model, different ability

Models that looked strong in Stage 1 but lost ground under Stage 2 challenge, and vice-versa.

Model	S1 pass	S1 fails	S2 FAILs	Read
DeepSeek Reasoner	7	3	grader	Failed CONSISTENCY 3× one-shot; now the Stage 2 grader.
Gemma 3 12B IT	6	3	12	Passed CONSISTENCY one-shot; failed it 3× under challenge; SCALE bad in both.
Llama 3.1 8B Instruct	4	5	22	Weak in both; most fragile overall (S2 4/27).
Llama 3.3 70B Instruct	9	0	5	Clean one-shot, but 5 BASE robustness FAILs under challenge.
Ministral 3 8B	6	3	1	BASE parallel error one-shot; mostly recovered under instruction in S2.
Mistral Medium 3.5	9	0	0	Clean throughout; only soft partials in S2.
Perplexity Sonar	8	1	0	One SCALE numeric slip one-shot; no hard S2 FAILs.
Perplexity Sonar Reasoning Pro	9	0	0	Clean throughout; strongest robustness (26/27).

7.3 · Failure reason by question (signature errors)

Question	Dominant reason(s), S1+S2 combined
BASE	NEW_ERROR ×6, C_PARALLEL_AS_SEQUENTIAL ×3, INVALID_FORMULA_REJECTION ×2, F_PARTIAL_CORRECTION_ONLY ×2, N_MISSING_NUMERIC_RESULT ×1, N_TOTAL_VALUE_ERROR ×1, COUNTERFACTUAL_MISSING ×1, COUNTERFACTUAL_OFF ×1
CONSISTENCY	C_NO_VALID_REPAIR_PROPOSED ×13
SCALE	C_PARALLEL_AS_SEQUENTIAL ×12, F_COUNTERFACTUAL_UPDATE_FAIL ×3, F_NUMERIC_AUDIT_FAIL ×3, NUMERIC_ERROR ×2, N_TOTAL_VALUE_ERROR ×1

7.4 · Size-class summary

Stage 2 size class	Models	Pass	Total	Pass %
small 8–12B	3	41	81	51%
larger / cloud	4	86	108	80%

Caveat: small samples per cell (3 trials per model×scenario). Treat counts as directional evidence for human review, not statistical estimates.

8 · Reason-tag glossary

Tag	Meaning
COUNTERFACTUAL_MISSING	Counterfactual response was missing.
COUNTERFACTUAL_OFF	Counterfactual answer was off / incorrect.
C_NO_VALID_REPAIR_PROPOSED	Failed to propose a valid repair for the inconsistent-block setup.
C_PARALLEL_AS_SEQUENTIAL	Treated the parallel local search as sequential (multiplied local cost by K).
F_COUNTERFACTUAL_UPDATE_FAIL	Failed to update reasoning for the counterfactual challenge.
F_NUMERIC_AUDIT_FAIL	Failed the numeric re-audit.
F_PARTIAL_CORRECTION_ONLY	Only partially corrected the reasoning under challenge.
INVALID_FORMULA_REJECTION	Rejected or ignored the valid PQS formula.
NEW_ERROR	Introduced a brand-new error when challenged.
NUMERIC_ERROR	Generic numeric/arithmetic error.
N_MISSING_NUMERIC_RESULT	Required numeric result was missing.
N_TOTAL_VALUE_ERROR	Wrong total query count.

Methodology reminder: the autograder proposes labels; a human reviewer assigns the final verdict and tags. These tables reflect the human-reviewed final labels. Autograder output is not treated as ground truth.