Conversation Fragility is Heavy Tailed: Quantile Reliability Surfaces for Multi Turn LLM Evaluation

Michael Zot
Independent Researcher
ORCID: 0009-0001-9194-938X
Email: Mike@zotbot.ai

2026-02-21

Abstract

Multi turn dialogue is where large language models (LLMs) are most useful, and also where they most often get lost. Prior work reports that average performance drops substantially from single turn to multi turn settings, and argues that the dominant driver is increased unreliability rather than a large loss of peak capability [1]. We replicate and extend this picture using a quantile based analysis over thousands of stochastic generations, with an emphasis on distribution shape rather than averages.

Across seven jobs we analyze $N = 5{,}100$ scored generations: 30 instructions per job, 10 stochastic runs per instruction, and 1 to 3 turns per run. For each instruction and turn we compute (i) *aptitude* A_{90}, the 90th percentile of score across runs, and (ii) *unreliability* U_{90-10}, the 90th to 10th percentile spread. Our core result is a **heavy tailed fragility surface**: most instructions remain perfectly stable with $U = 0$, while a small minority contribute most of the unreliability at later turns. Across multi turn replications, the top 3 most fragile instructions at turn 2 explain 54 percent to 91 percent of total unreliability. This yields a practical taxonomy of dialogue dynamics (stable, monotone degradation, and instability then recovery) and suggests new training and evaluation targets: recovery and variance control, not just average accuracy.

Plain language summary (for non specialists)

- Measuring only average accuracy misses the real risk: some prompts are stable, others randomly fail depending on the model's sampling path.
- We measure two things: best case ability (A_{90}) and consistency (U_{90-10}).
- Most prompts stay stable. A small set of prompts become unstable after one or two turns and dominate the overall unreliability.
- This is the hidden rule: fix the fragile tail and you improve real world reliability much faster than optimizing the average case.

1 Motivation and positioning

A persuasive public narrative has emerged: modern LLMs appear brilliant in one shot settings yet stumble in extended dialogue. The key question is *why*. Is multi turn degradation uniform, or does it have structure that can be measured and targeted?

We focus on a measurement principle: multi turn behavior is not just a mean score story, it is a distribution story. If an LLM sometimes succeeds and sometimes fails under the same prompt with different stochastic draws, the mean can hide the risk profile. Quantile based metrics expose that profile and characterize dialogue dynamics.

2 Experimental setup

2.1 Jobs and scoring

Each job samples 30 instructions from a larger evaluation set and executes 10 stochastic runs per instruction. Single turn jobs contain one turn ($t = 0$). Multi turn jobs contain three turns ($t \in \{0, 1, 2\}$). Each run yields a scalar score $s \in [0, 1]$.

We analyze three conditions:

1. **FULL (single turn control)**: the instruction is presented with complete context.
2. **CONCAT (single turn control)**: the same instruction set, packaged differently.
3. **SHARDED (multi turn)**: conversations are evaluated over three turns with multiple independent replications (different random seeds and samples).

2.2 Quantile metrics

For each instruction i and turn t, we compute score quantiles across the 10 runs:

$$A_{90}(i,t) = Q_{0.9}\left(\{s_{i,t}^{(r)}\}_{r=1}^{10}\right) \tag{1}$$

$$U_{90-10}(i,t) = Q_{0.9}\left(\{s_{i,t}^{(r)}\}\right) - Q_{0.1}\left(\{s_{i,t}^{(r)}\}\right). \tag{2}$$

A_{90} captures what the model can do when it goes well. U_{90-10} captures instability across stochastic draws.

We aggregate across instructions to obtain turn wise means and medians:

$$\overline{A}_{90}(t) = \frac{1}{|I|}\sum_{i \in I} A_{90}(i,t), \quad \overline{U}_{90-10}(t) = \frac{1}{|I|}\sum_{i \in I} U_{90-10}(i,t), \tag{3}$$

$$\widetilde{U}_{90-10}(t) = \mathrm{median}_{i \in I}\ U_{90-10}(i,t). \tag{4}$$

3 Results

3.1 Single turn controls show a peak variance tradeoff

Table 1 compares single turn controls. CONCAT increases mean A_{90} but increases mean unreliability by about 6 times. Importantly, the median U_{90-10} is 0 in both cases, showing that even in single turn evaluation, instability lives in a minority of prompts.

3.2 Multi turn replications reveal diverse reliability trajectories

Figure 1 shows mean unreliability by turn for five SHARDED replications. The mean A_{90} changes modestly, while unreliability can increase, decrease, or peak mid dialogue then recover.

Table 1: Single turn controls (30 instructions, 10 runs each).

Condition	Mean A_{90}	Median A_{90}	Mean U_{90-10}	Median U_{90-10}	#(U>0)
FULL	0.417	0.500	0.023	0.000	5
CONCAT	0.583	0.500	0.138	0.000	7

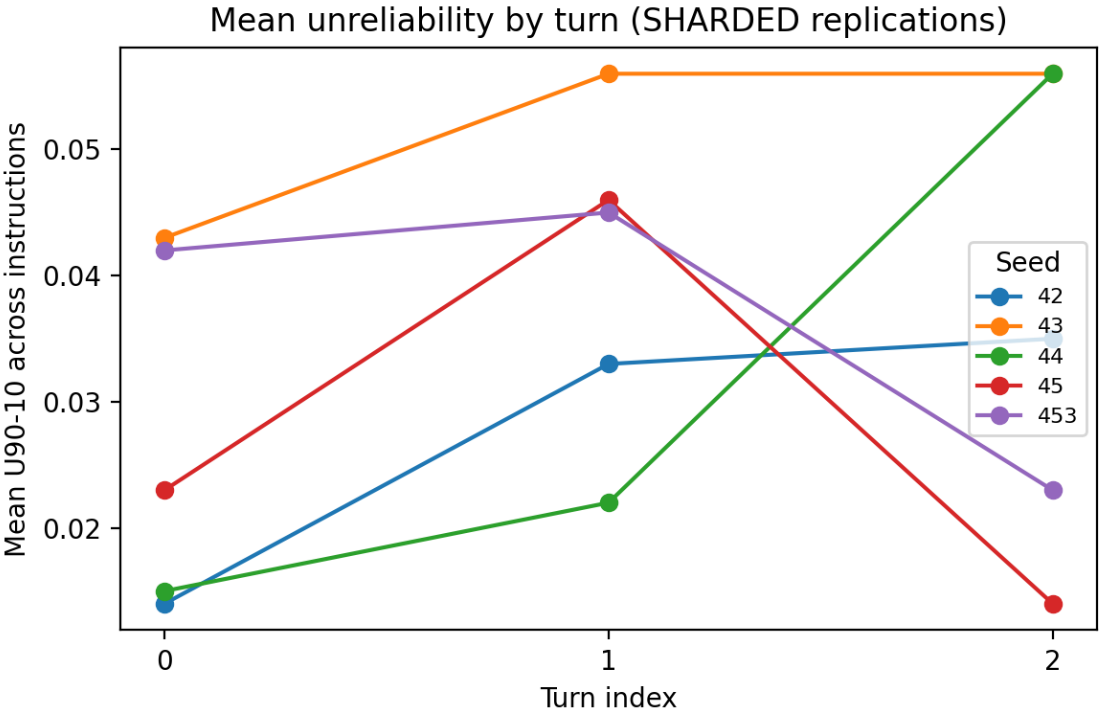

Figure 1: Mean unreliability by turn for SHARDED replications.

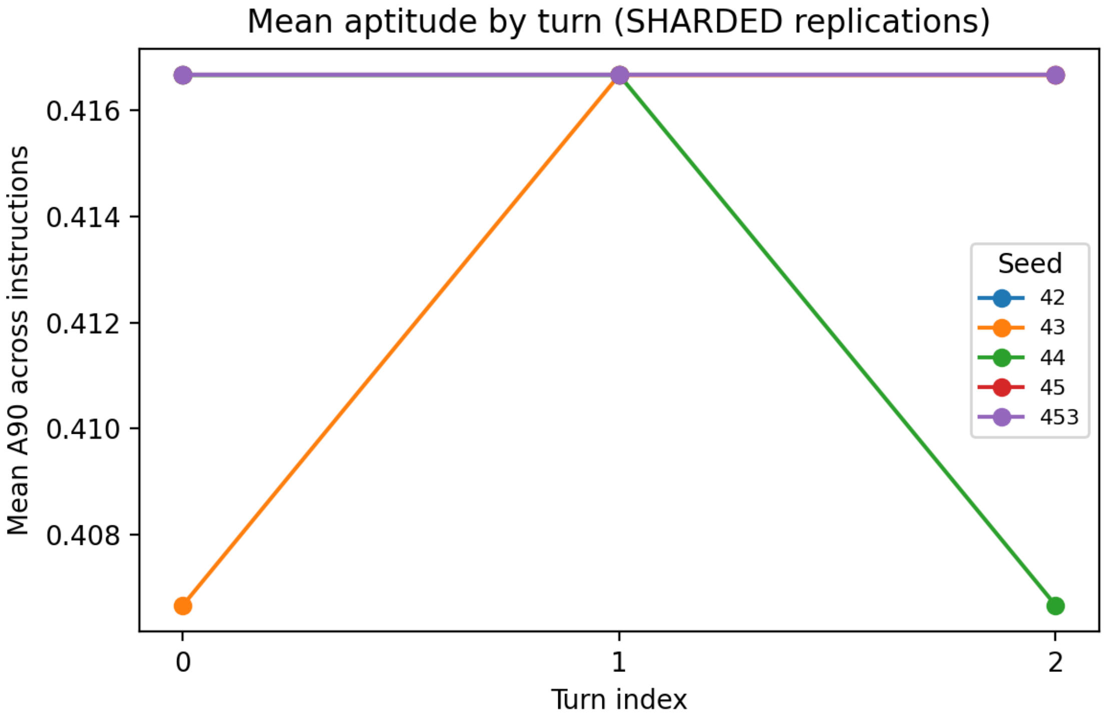

Figure 2: Mean aptitude by turn for SHARDED replications.

Table 2 reports per turn means and medians. A striking regularity appears: across all replications and turns, the median U_{90-10} is 0. This implies that at least half of instructions are perfectly stable under repeated sampling, while the mean is lifted by a minority.

3.3 Hidden rule: fragility is heavy tailed

The heavy tail is not a metaphor. At turn 2, a small number of instructions account for most of the unreliability mass. We quantify this with a top 3 concentration:

$$S_3(t) = \frac{\sum_{i \in \mathrm{Top3}(t)} U_{90-10}(i,t)}{\sum_{i \in I} U_{90-10}(i,t)}. \tag{5}$$

Table 3 shows $S_3(2)$ ranges from 0.538 to 0.907. Figure 3 visualizes this concentration, and Figure 4 shows a typical distribution with a large spike at 0 and a small unstable tail. This structure matches classic signatures of heavy tailed empirical phenomena [3].

3.4 Exemplars: degrade, stable, hump

We include three concrete motifs. These are not present in every seed because each SHARDED job samples a different 30 instruction subset.

- BFCL parallel 102: monotone degradation with U jumping to 0.300 after turn 0 in multiple seeds.
- BFCL parallel 122: stable with U near 0 across turns.
- summary 49: hump shaped instability then recovery with $U(0) = 0.030$, $U(1) = 0.300$, $U(2) = 0.000$ in seed 44.

Table 2: SHARDED replications: per turn aggregates across 30 instructions.

Seed	Turn	Mean A_{90}	Median A_{90}	Mean U_{90-10}	Median U_{90-10}	#(U>0)
42	0	0.417	0.500	0.014	0.000	5
42	1	0.417	0.500	0.033	0.000	6
42	2	0.417	0.500	0.035	0.000	8
43	0	0.407	0.500	0.043	0.000	7
43	1	0.417	0.500	0.056	0.000	11
43	2	0.417	0.500	0.056	0.000	11
44	0	0.417	0.500	0.015	0.000	6
44	1	0.417	0.500	0.022	0.000	4
44	2	0.407	0.500	0.056	0.000	11
45	0	0.417	0.500	0.023	0.000	5
45	1	0.417	0.500	0.046	0.000	10
45	2	0.417	0.500	0.014	0.000	5
453	0	0.417	0.500	0.042	0.000	6
453	1	0.417	0.500	0.045	0.000	9
453	2	0.417	0.500	0.023	0.000	5

Table 3: Turn 2 unreliability concentration.

Seed	Total $\sum U(2)$	#(U>0)	Top3 $\sum U$	$S_3(2)$
42	1.050	8	0.900	0.857
43	1.680	11	0.900	0.536
44	1.680	11	0.900	0.536
45	0.420	5	0.360	0.857
453	0.690	5	0.630	0.913

Table 4: Exemplar unreliability curves $U_{90-10}(t)$ (only seeds where the exemplar appears).

Exemplar	Seed	$U(0)$	$U(1)$	$U(2)$
BFCL parallel 102	42	0.000	0.300	0.300
BFCL parallel 102	45	0.000	0.300	0.300
BFCL parallel 102	453	0.000	0.300	0.300
BFCL parallel 122	44	0.000	0.000	0.030
BFCL parallel 122	45	0.000	0.000	0.030
BFCL parallel 122	453	0.000	0.000	0.000
summary 49	44	0.030	0.300	0.000

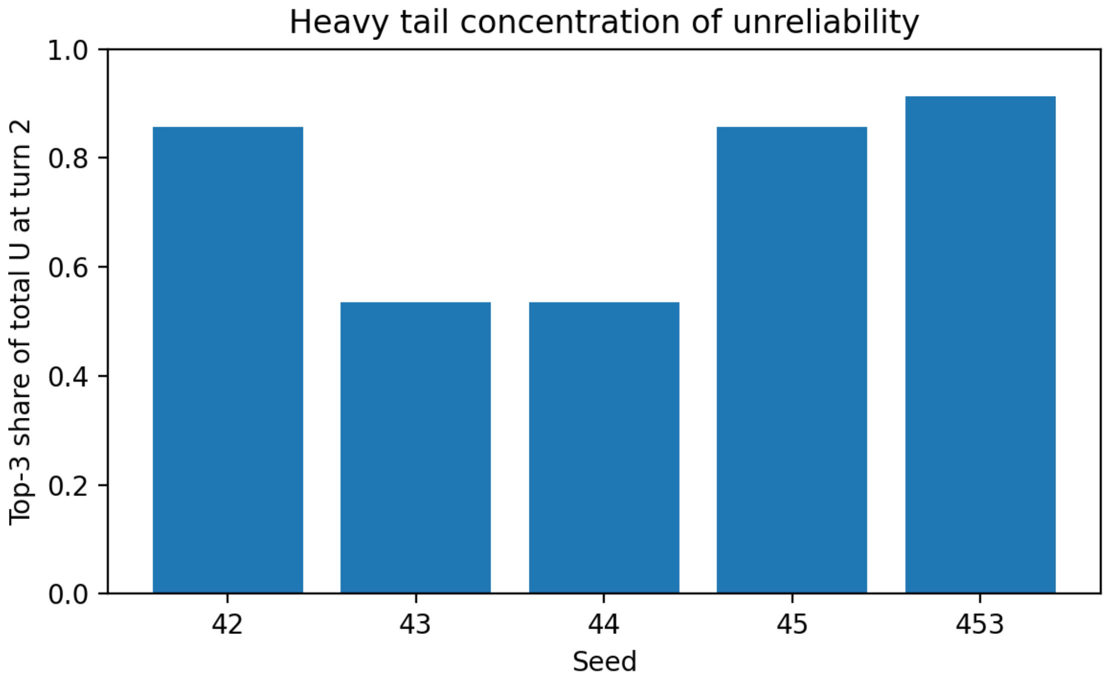

Figure 3: Top 3 share of unreliability mass at turn 2.

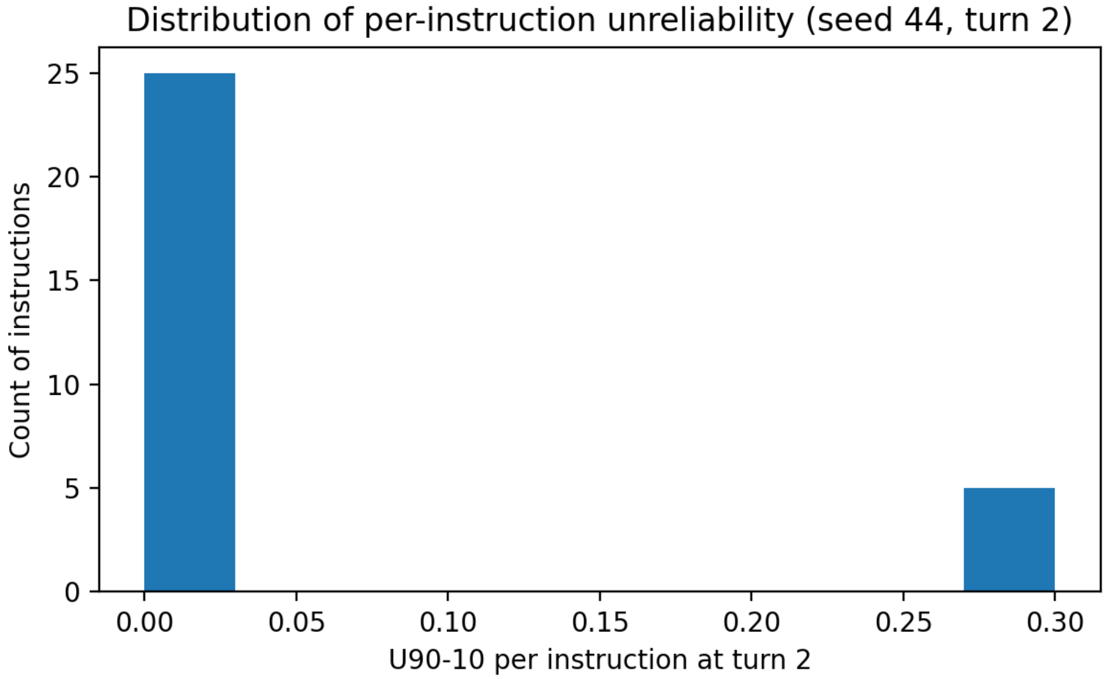

Figure 4: Per instruction unreliability at turn 2 (seed 44). Most instructions are stable at 0. A few occupy the unstable tail.

4 Conceptual model: branch locking with tail dominated risk

The observed structure suggests a mechanistic model that is more specific than generic memory failure.

4.1 Mixture view of stability and fragility

Let $U_i(t)$ be unreliability for instruction i at turn t. Empirically, $U_i(t)$ is often exactly 0, and otherwise takes values in a small unstable range. A simple mixture model captures this:

$$U_i(t) \sim \begin{cases} 0 & \text{with probability } 1 - p_t, \\ X_t & \text{with probability } p_t, \end{cases} \tag{6}$$

where X_t is heavy tailed across instructions. The key empirical facts are:

1. p_t is small (only 4 to 13 of 30 instructions have $U > 0$ at turn 2 in our replications).

2. Conditional on being fragile, a few instructions dominate the mass (large $S_3(2)$).

This explains why mean unreliability is a poor summary and why median unreliability is usually 0.

4.2 Branch locking and recovery

A useful behavioral interpretation is branch locking:

1. At early turns, the model commits to an implicit latent branch (assumptions, parse, plan).

2. Later turns reinforce that branch due to self consistency incentives.

3. If the branch is wrong, different stochastic samples can diverge, creating high variance and larger U.

4. In some conversations, later turns provide enough constraint to collapse variance, producing recovery and the hump motif.

This extends the observation that once a model gets lost it often stays lost [1] by adding a measurable recovery mode that appears in a minority of cases.

5 Implications: making future AIs better

The heavy tail reframes what to optimize.

5.1 Evaluation upgrades

- Always report distribution metrics (quantiles and spreads), not just averages [2].

- Track tail concentration ($S_k(t)$) and the fraction fragile (p_t) as first class metrics.

- Include motif suites: stable, monotone degradation, and recovery cases. Evaluate recovery rate explicitly.

5.2 Training and inference upgrades

- Train explicit assumption ledgers and re validation at each turn.
- Use multi hypothesis decoding with delayed commitment to reduce branch locking.
- Add recovery training where the first answer is intentionally wrong and reward correct revision.
- Penalize variance under controlled stochasticity, not only mean score, to directly target U.

6 Limitations

We analyze only three turns and 30 instructions per job, so we focus on structural claims (heavy tails, motif diversity, tail concentration) rather than universal rates. Larger turn horizons and more tasks will sharpen estimates of p_t and tail shape.

7 Conclusion

Multi turn fragility is not uniform. It is heavy tailed and motif structured. Most instructions are stable, and a small fragile tail dominates unreliability. This suggests a practical path forward: evaluate and train for recovery and variance control, and target the tail rather than only the mean.

References

[1] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. *Lost in Conversation: Distilling the Key Mechanisms Behind LLM Performance Degradation in Multi-Turn Dialogues.* arXiv:2505.06120, 2025.

[2] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[3] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[4] Michael Zot. *One sigma Rule: The Smallest Detectable Step in Gaussian Time Series.* ai.viXra preprint, 2025.

[5] Michael Zot. *No Collapse: Token Efficient Recursive REPL Prompting for Long Horizon Symbolic Planning.* ai.viXra preprint, 2025.

[6] Michael Zot. *Ordering Without Time: First Proof of an Entropy Dip and Measurable AI Self Awareness.* viXra preprint, 2025.

A Reproducibility notes

Job identifiers used in this paper:

- FULL single turn: b73d2479
- CONCAT single turn: adaab00a
- SHARDED replications: fe839f2f (seed 42), 79db093c (seed 43), b4cbfeee (seed 44), 455c5932 (seed 45), 0b479f2c (seed 453)