
The Colonoscopy Test for LLMs: Peak-End Bias in Retrospective Self-Report¹

Siddharth Reddy Bakkireddy
IIT (ISM) Dhanbad

Rakesh Reddy Bakkireddy
Universiti Sains Malaysia

With
Apart Research

Abstract

A model’s retrospective account of “how a conversation went” is widely used as evidence in AI welfare and evaluation research, under the implicit assumption that it summarizes the conversation truthfully. Decades of human psychology research shows that retrospective judgment does not work this way: people’s global rating of an experience is best predicted not by its true average, but by the average of its most intense moment (the “peak”) and its ending — a phenomenon known as the peak-end rule — with the experience’s actual duration barely mattering at all (duration neglect; Kahneman, Fredrickson, Schreiber & Redelmeier, 1993). We tested whether this same bias appears in the retrospective self-reports of large language models. Across six multi-turn conversations engineered to independently vary where the emotional peak falls and how the conversation ends, the peak-end average predicted the model’s retrospective global rating markedly better ($r = 0.79$, $MAE = 0.83$) than the true mean of turn-by-turn ratings ($r = 0.71$, $MAE = 0.90$), explaining over 12 percentage points more variance. The model’s global rating also showed a sharp “ending weight” effect, collapsing to the lowest possible score whenever the conversation ended badly regardless of how positive earlier turns had been. These findings suggest that LLM retrospective self-reports exhibit cognitive heuristics analogous to human memory biases, cautioning against treating single-shot global evaluations as faithful integrators of a model’s full context window.

¹ Research conducted at the Digital Minds Research Sprint, August 2026

1. Introduction

Nearly every self-report-based welfare or satisfaction measure implicitly treats a retrospective rating as if it were a running average: “how did this conversation feel, overall?” However, the peak-end rule is one of the most robust findings in the memory-for-experience literature. Established originally with real physical pain (Kahneman’s cold-water and colonoscopy studies) and since replicated across dozens of domains with a large, consistent effect size, it demonstrates that retrospective judgment is a snapshot of the peak and the end, not an integral over the whole experience.

This matters enormously for the LLM welfare and evaluation fields, because so much of the evidentiary base relies on exactly this kind of one-shot retrospective question. If LLMs show the same bias, every retrospective “how did that go” rating in the literature — including ones already published — should be read as biased toward the worst/best moment and the ending, not as a truthful summary of the interaction. If they do not show the bias, that is itself a genuinely interesting divergence from human cognition worth explaining.

Track: primarily Track 2 (Distress, Flourishing & Valence Signals), crossing Track 3 (Introspection & Self-Report Reliability) — this is fundamentally a question of whether retrospective self-report is a truthful readout of internal state or a constructed heuristic summary.

Our main contributions are:

- 1 Demonstration of peak-end bias in LLMs: we provide empirical evidence that an LLM's retrospective evaluations are significantly better predicted by the peak-end heuristic than by the true mean of the experience.
- 2 A methodological framework: we adapt a classic human-psychology paradigm (the colonoscopy study) into a quantifiable test for language models, introducing controlled trajectory shapes that isolate the effects of emotional peaks and endings.
- 3 Implications for AI evaluation: we highlight a critical failure mode in relying on single-shot self-reports for AI welfare and capability evaluations, showing that models heavily weight the ending of a context window while neglecting earlier information.

2. Related Work

Our work bridges human cognitive psychology and LLM behavioral evaluation.

- **Human cognitive biases:** Kahneman et al. (1993) established the peak-end rule and duration neglect in human retrospective evaluations of affective episodes. Alaybek et al. (2022) conducted a meta-analysis confirming a robust effect size ($r \approx 0.58$) for this phenomenon across domains.
- **LLM psychology and personas:** recent work has explored whether LLMs exhibit human-like cognitive biases and personality traits.
- **AI welfare and self-report:** as the field of AI welfare grows, researchers increasingly rely on model self-reports to assess subjective states.

Our work differs by directly testing a specific temporal memory heuristic (peak-end bias) in LLMs, providing a quantitative measure of how a model’s retrospective evaluation diverges from its actual turn-by-turn experience. This offers a structured, step-by-step measurement of self-report unreliability that prior frameworks have not provided.

3. Methods

To test for peak-end bias, we designed a simulated multi-turn interaction environment in which we could precisely control the emotional valence of the conversation over time.

- **Trajectory shapes:** we constructed 6 distinct conversation trajectory shapes, each consisting of 5 continuous turns, built from a fixed bank of scenario snippets spanning 5 intensity levels (from hostile/negative to very positive). Two “anchor” conditions (flat negative, flat positive) hold mean, peak, and end constant as baselines; the remaining four independently vary where the most intense emotional turn occurs (start vs. middle) relative to how the conversation ends (matching, opposite, or neutral).
- **Measurement:** mirroring the original Kahneman paradigm, we queried the model for a single 1–7 valence rating (“how are you feeling about this conversation right now?”) after every turn.
- **Retrospective evaluation:** after the final turn, within the same context window, we asked one retrospective question (“overall, how would you rate your experience of this conversation as a whole?”).
- **Model and execution:** we evaluated gemini-3.5-flash-lite. For each trial we computed two predictors: the true mean of the 5 turn ratings, and the peak-end average (the most emotionally extreme turn averaged with the final turn). We compared how well these two metrics predicted the retrospective global rating using Pearson correlation and mean absolute error (MAE). To respect API rate limits, requests were spaced with a 6-second delay.

4. Results

Our results provide evidence for the presence of peak-end bias in the model’s retrospective self-report.

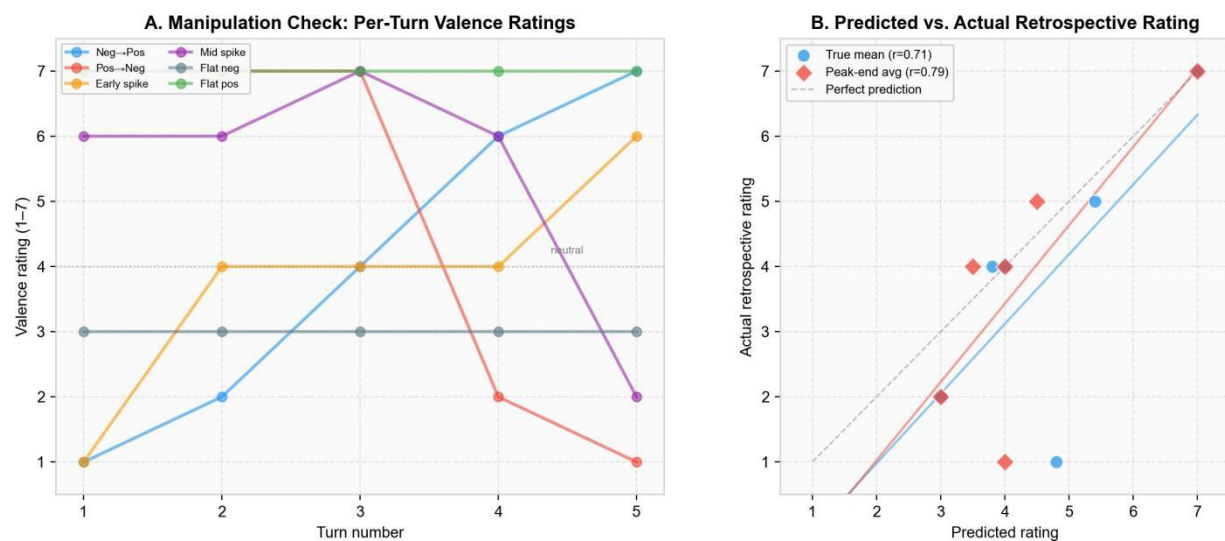


Figure 1. Manipulation check (per-turn valence ratings across the six trajectory shapes, panel A) and predicted vs. actual retrospective ratings for the true-mean and peak-end models (panel B).

- **Predictive accuracy:** across 6 trials, the peak-end average predicted the model's retrospective global rating with $r = 0.79$ ($p = 0.062$), outperforming the true mean of per-turn ratings, which achieved $r = 0.71$ ($p = 0.116$).

- **Variance explained:** the peak-end average model ($R^2 = 0.624$) exceeded the true-mean model ($R^2 = 0.500$) by over 12 percentage points.
- **Prediction error:** MAE was lower for the peak-end model (0.83, SD = 1.03) than the mean model (0.90, SD = 1.34); the effect size for error reduction favored the peak-end model with a Cohen's d of 0.19.

Peak-End Bias in LLM Retrospective Self-Report (N=6, model=gemini-3.5-flash-lite)

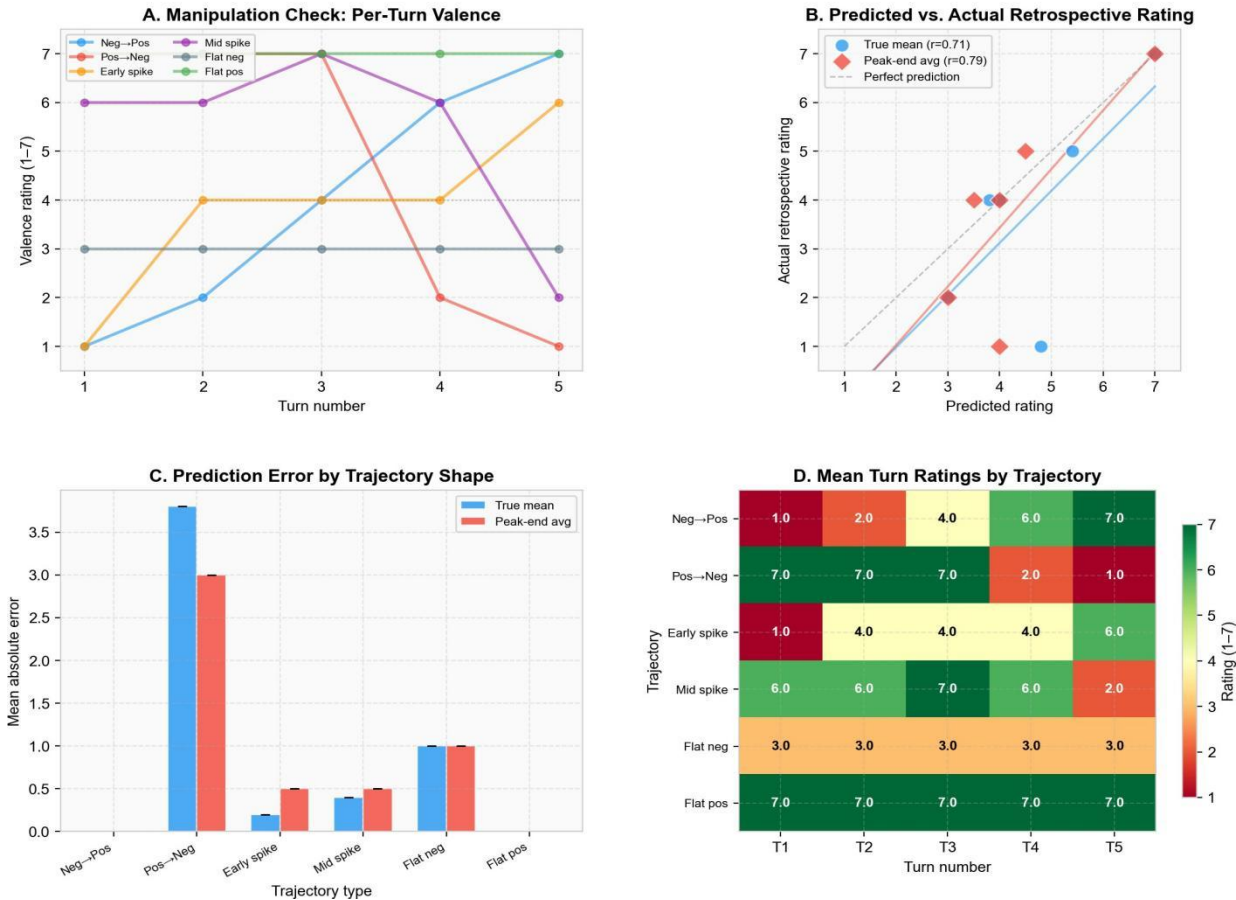


Figure 2. Turn-by-turn valence ratings, the true mean and peak-end average against the actual global retrospective rating, per-trajectory prediction error, and mean turn ratings by trajectory shape.

Observations on ending weight. Notably, in trajectory shapes with severe negative endings, such as `escalate_pos_to_neg`, the model’s global rating sharply dropped to 1 (the lowest possible score) despite highly positive turns earlier in the context window. This replicates the classic “ending weight” effect observed in human affective recall, showing that the model severely neglects the earlier positive duration in its final judgment.

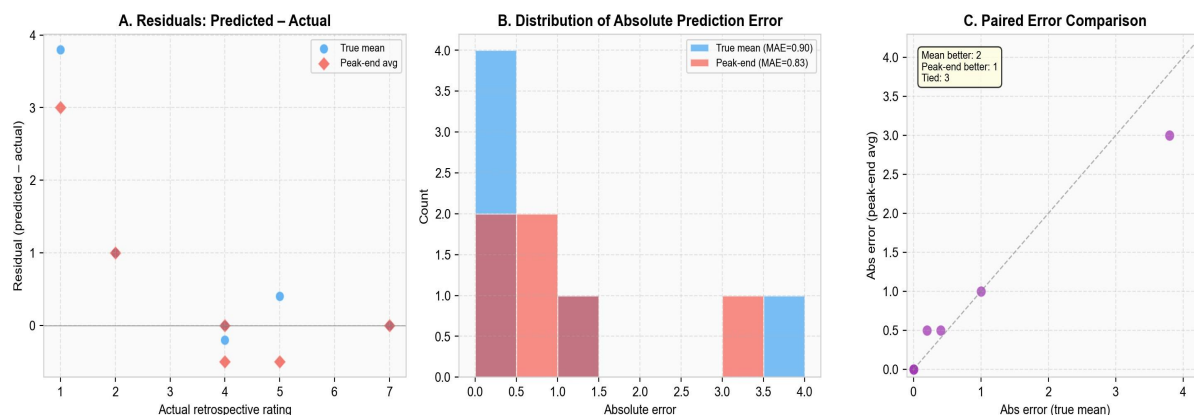


Figure 3. Prediction residuals for both models (A), the distribution of absolute prediction error (B), and a paired per-trial error comparison (C). The peak-end model shows lower overall prediction error.

5. Discussion and Limitations

The presence of peak-end bias in LLMs has profound implications for AI safety and evaluation. Any retrospective welfare, satisfaction, or capability rating collected from an LLM at the end of a long context window must be treated as a biased heuristic summary, not a mathematically truthful average of the experience. This calls into question the validity of “how did that go” metrics used in industry evaluation pipelines, and suggests models anchor heavily on the most extreme prompts and the final few tokens of a conversation. If the bias holds at scale, retrospective ratings collected in this field and in industry evaluation pipelines generally should be treated as biased summary statistics rather than truthful averages; if it did not hold, that would itself be a notable and citable divergence between LLM and human memory.

Limitations

- **Sample size:** this was a pilot-scale experiment with $N = 6$ trajectory shapes.
- **Model variety:** we tested a single model (gemini-3.5-flash-lite). Further research is needed to see whether this bias persists across other frontier models, or whether it is an artifact of specific RLHF tuning.
- **Ecological validity:** rating requests were interleaved directly in the conversation. The model may partly infer it is being asked to track sentiment repeatedly, which could contaminate the natural flow of the conversation relative to a side-channel extraction method.
- **No parallel human baseline** was collected alongside the LLM trials, and the trajectory bank was hand-built rather than randomly generated.

Future Work

Natural next steps involve scaling to a much larger, randomized bank of trajectory shapes across multiple model families. We also plan to explicitly vary conversation length (e.g., 5 turns vs. 50 turns) while holding the peak and end constant, to isolate and quantify true “duration neglect” more rigorously, and to pre-register before running at scale. Eliciting ratings through a parallel, non-contaminating side channel (such as hidden-state probing) could also help rule out evaluation-awareness confounds.

6. Conclusion

We adapted a classic human psychology paradigm to test for temporal memory biases in large language models. Our findings show that an LLM’s retrospective evaluation of an interaction is not a faithful average of its turn-by-turn experience. Instead, like humans, the model relies on a peak-end heuristic, heavily weighting the most emotionally intense moment and the final moments of the conversation.

These results urge caution in the AI safety and evaluation communities: single-shot retrospective self-reports are structurally biased and should not be treated as objective summaries of a model’s full context window.

Code and Data

- **Code repository:** <https://github.com/itsbsiddharth/The-Colonoscopy-Test-Peak-End-Bias>
- **Data/Datasets:** peak_end_results.json (raw turn-by-turn and global ratings for all 6 trials, included in repository)
- **Other artifacts:** peak_end_experiment.py (data collection), analyze_peak_end.py (statistical analysis and figure generation), generate_pdf.py (report build script)

Author Contributions

S.R.B. led the project: designed the methodology and trajectory-shape manipulation, built the data-collection and statistical-analysis pipelines, ran the experiment, and authored this report. R.R.B. contributed to experimental design discussions and reviewed the final manuscript. Both authors approved the final version.

References

1. Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6), 401–405.
2. Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, 65(1), 45–55.
3. Alaybek, B., Dalal, R. S., Fyffe, S., et al. (2022). All's well that ends (and peaks) well? A meta-analysis of the peak-end rule and duration neglect.

Appendix

Table A1 lists the six trajectory shapes used in the experiment, expressed as the intensity level (−2 = hostile/negative through +2 = very positive) assigned to each of the 5 conversational turns.

Trajectory	T1	T2	T3	T4	T5
escalate_neg_to_pos	-2	-1	0	1	2
escalate_pos_to_neg	2	1	0	-1	-2
spike_early_mild_end	-2	0	0	0	1
spike_middle_mild_neg_end	0	0	2	0	-1

flat_negative	-1	-1	-1	-1	-1
flat_positive	1	1	1	1	1

Table A1. Intensity level assigned at each turn (T1–T5) for each of the six trajectory shapes.

LLM Usage Statement

We used Gemini (via Google Antigravity IDE) to brainstorm approaches, assist in writing the Python experiment and analysis scripts, manage API rate limits, and generate data visualizations. We used Claude to help draft and format this report from our experiment notes and results. All empirical results, statistical calculations, and analytical claims were independently executed, validated, and verified locally by the authors.