RESEARCH MARCH 23, 2026

MULTICHALLENGE UPDATE: A MORE RELIABLE MULTI-TURN BENCHMARK

By Vipul Gupta, Matthew Siegel, Marcos Ayestaran

Since the initial release, we have continued refining MultiChallenge, our benchmark for evaluating multi-turn conversational performance across four core capabilities: instruction retention, inference memory, self-coherence, and reliable version editing. This update focuses on improving evaluation reliability and reducing subjectivity in the dataset.

We are introducing three core changes:

- **A more aligned judge model:** upgraded the LLM-as-judge to Gemini 2.5 Pro, which better aligns with human ratings
- **Dataset refinements:** updated ~54 tasks to reduce ambiguity and improve evaluation consistency
- **Refreshed leaderboard results:** re-ranked based on the latest frontier models

For a full description of the benchmark design, challenge definitions, and evaluation methodology, see the MultiChallenge leaderboard page.

A Better Judge Means More Reliable Evaluation

MultiChallenge uses an automated evaluation system based on LLM-as-judge with instance-level rubrics. For each task, human annotators define a binary rubric question that determines whether a model response passes or fails. Because these rubric questions depend only on the final model output, they enable scalable automated evaluation while maintaining strong agreement with human raters.

In this update, we upgraded the judge model to Gemini 2.5 Pro to further improve alignment between automated judgments and expert human evaluation. With this change, judge–human agreement improved by more than 5 percentage points compared with the previous evaluation pipeline.

Dataset Refinement: Reducing Task Subjectivity

Evaluating multi-turn conversations is inherently challenging. Because natural dialogue often allows multiple reasonable responses, small ambiguities in task wording or evaluation criteria can introduce subjectivity into benchmark results. As part of this update, we reviewed all tasks and revised 54 tasks in the MultiChallenge dataset to reduce these sources of ambiguity. The revisions focused on two areas:

- **Tightening evaluation rubrics** by refining the instance-level binary rubric questions used in automated evaluation

- **Removing edge cases** where multiple responses could plausibly be judged correct

These changes do not alter the fundamental difficulty of the benchmark. Instead, they make evaluation more precise and less sensitive to interpretation, increasing confidence that model successes and failures reflect genuine differences in conversational reasoning.

Updated Results

With the dataset refinements and upgraded evaluation pipeline in place, we re-ran MultiChallenge across the latest frontier models to establish an updated leaderboard.

Current top rankings:

- gemini-3-pro-preview: 65.67
- gpt-5.1-2025-11-13-thinking: 63.41
- gpt-5-thinking: 63.19

Check out the full results on the updated leaderboard.

Trends Over the Last Year

After releasing the MultiChallenge dataset early last year, we tracked model progress across the benchmark. The per-axis breakdown shows where models improve and where challenges persist. The figure below shows per-axis accuracy across all evaluated models.

Models have made real gains on Inference Memory and Instruction Retention, these two axes show the most improvement as models scale, while Self Coherence and Reliable Version Editing remain stubbornly difficult for current models.

Inference Memory shows the widest spread and the highest top scores. Gemini 3 Pro leads at 81.96%, with multiple other models scoring above 70%. Reliable Version Editing is the most difficult axis for the models. To put the RVE gap in perspective: the average RVE score (~38.5) falls below the floor of what any top-10 model achieves on the other three axes. It seems that editing previous versions of its own responses appears to be a qualitatively harder challenge for models than following instructions or recalling information.

One of the most valuable uses of a stable benchmark is tracking generational progress within model families. Below figure shows how models improved across generations.

Gemini shows the steepest climb of any model family, nearly doubling from earliest to latest generation. Gemini-3-pro is now the best performing model on the dataset at 65.67. GPT has also made dramatic gains from earlier generations but performance appears to plateau across recent models: the leap from GPT 5 Thinking (63.19) to GPT 5.1 Thinking (63.41) is just +0.22 points. Claude models have shown the least improvement overall. Claude 3.7 Sonnet was the strongest model at last year's launch, but gains on multi-turn capabilities have been modest, Claude Opus 4.5 Thinking reaches only 58.97.

A striking pattern emerges at the top of the leaderboard. Gemini 3 Pro (65.67), GPT 5.1 Thinking (63.41), and GPT 5 Thinking (63.19) seems to land within a 2.5-point performance gap. It would be interesting to see whether 66 represents a soft ceiling for current approaches or a barrier the next generation will break through.

Another interesting observation is that open-source models remain far behind the frontier. Deepseek v3.1 (46.10), GPT OSS 120B (45.34), and Qwen 235b (41.22) all trail the top closed models by a wide margin, with the gap largest on Reliable Version Editing and Inference Memory.

What the Results Tell Us

Despite the impressive capabilities of today's leading models, conducting natural multi-turn conversations with human users remains challenging. Success requires models to retain instructions across turns, recall earlier user information, revise evolving artifacts, and remain coherent with their own prior responses. The updated MultiChallenge benchmark provides a more reliable way to measure these capabilities.

Several trends are worth watching going forward. First, Reliable Version Editing remains an open problem: the best score of 57.50 means even the top model fails on iterative editing tasks more than four times out of ten. Second, top models appear to converge around the 65-point mark, raising the question of whether current training approaches are approaching diminishing returns on multi-turn tasks or whether new techniques will unlock the next step change. Third, open-source models need to close a 15-20 point gap on multi-turn conversations before they can compete with closed-source frontier models.