



Validating LLM-based Evaluators Against Real User Ratings In A Conversational AI System For Cognitive Impairment

Esra Al, Ph.D., Daniel Kelly, Babak Parviz, Ph.D.
NewDays.ai
1201 2nd Ave, Suit 900, Seattle, WA 98101 USA

Introduction:

Cognition-oriented interventions are efficacious for individuals with dementia and mild cognitive impairment^{1,2}, and conversational AI offers a means of scaling their delivery. Large-scale deployment, however, requires continuous monitoring of conversational quality, which manual expert review cannot provide. Automated evaluation is therefore necessary, but only if its scores reflect the user's actual experience rather than superficial conversational features. We implemented an LLM-as-judge evaluator that scores the full call transcript, and assesses its concurrent validity against patient-reported satisfaction (CSAT), the most proximal available criterion for perceived call quality.

Methods:

We evaluated 729 calls from 36 patients (16 males; Mean age: 76.4, SD=6.4) in the NewDays cognitive intervention program, each with a post-call patient satisfaction rating (CSAT, 1–5) as ground truth. Three independently designed LLM-as-judge evaluators read the full call transcript and produced an automated quality score: a 5-point overall quality scale (overall-5), a 10-point quality scale that penalizes repetitive AI responses (overall-repeat-10), and a 100 point engagement scale (engagement-100). Each was run across four judge models (gpt-4o, gpt-4.1, claude-sonnet-4-6, gpt-5.5) and three temperatures (0.5–0.8) — a 17-configuration sweep — with three repeated samples averaged per call. We correlated each averaged score against CSAT using Pearson and repeated-measures correlations (the latter estimating the within-patient association), and tested for a length confound with partial correlations controlling for transcript words and user turns.

Evaluator	Model	Temp	Pearson r	95% CI	rm-r	p
overall-repeat-10	gpt-4o	0.7	0.383	[0.319, 0.443]	0.374	< 0.001
overall-repeat-10	gpt-4o	0.5	0.368	[0.303, 0.429]	0.362	< 0.001
engagement-100	gpt-4o	0.7	0.366	[0.302, 0.427]	0.332	< 0.001
engagement-100	gpt-4.1	0.7	0.363	[0.298, 0.424]	0.329	< 0.001
engagement-100	gpt-4.1	0.8	0.362	[0.298, 0.424]	0.324	< 0.001
overall-5	gpt-4o	0.7	0.361	[0.297, 0.423]	0.348	< 0.001
engagement-100	gpt-4o	0.8	0.355	[0.29, 0.417]	0.325	< 0.001
engagement-100	gpt-4o	0.5	0.353	[0.288, 0.415]	0.325	< 0.001
overall-repeat-10	gpt-4o	0.8	0.343	[0.277, 0.405]	0.339	< 0.001
overall-5	gpt-4o	0.8	0.338	[0.272, 0.401]	0.331	< 0.001
overall-5	gpt-4o	0.5	0.333	[0.267, 0.396]	0.309	< 0.001
overall-5	gpt-4.1	0.7	0.313	[0.246, 0.377]	0.302	< 0.001
overall-5	gpt-4.1	0.8	0.301	[0.234, 0.366]	0.290	< 0.001
overall-repeat-10	gpt-4.1	0.7	0.268	[0.199, 0.334]	0.260	< 0.001
overall-5	claude-sonnet-4-6	0.7	0.267	[0.198, 0.333]	0.232	< 0.001
overall-repeat-10	gpt-4.1	0.8	0.256	[0.187, 0.322]	0.249	< 0.001
overall-5	gpt-5.5	reasoning	0.213	[0.143, 0.282]	0.155	< 0.001

Table 1: Each of 17 LLM-judge configurations (evaluator × model × temperature) scored on the same 729 calls from 36 patients vs. patient satisfaction (CSAT, 1–5). Pearson r treats every call independently; rm-r is the within-patient correlation (the stricter validity test). All configs significant (p < 0.001); the near-identical r and rm-r show the signal is real call-to-call quality, not patient traits. Best config boxed: overall-repeat-10 · gpt-4o · t=0.7.

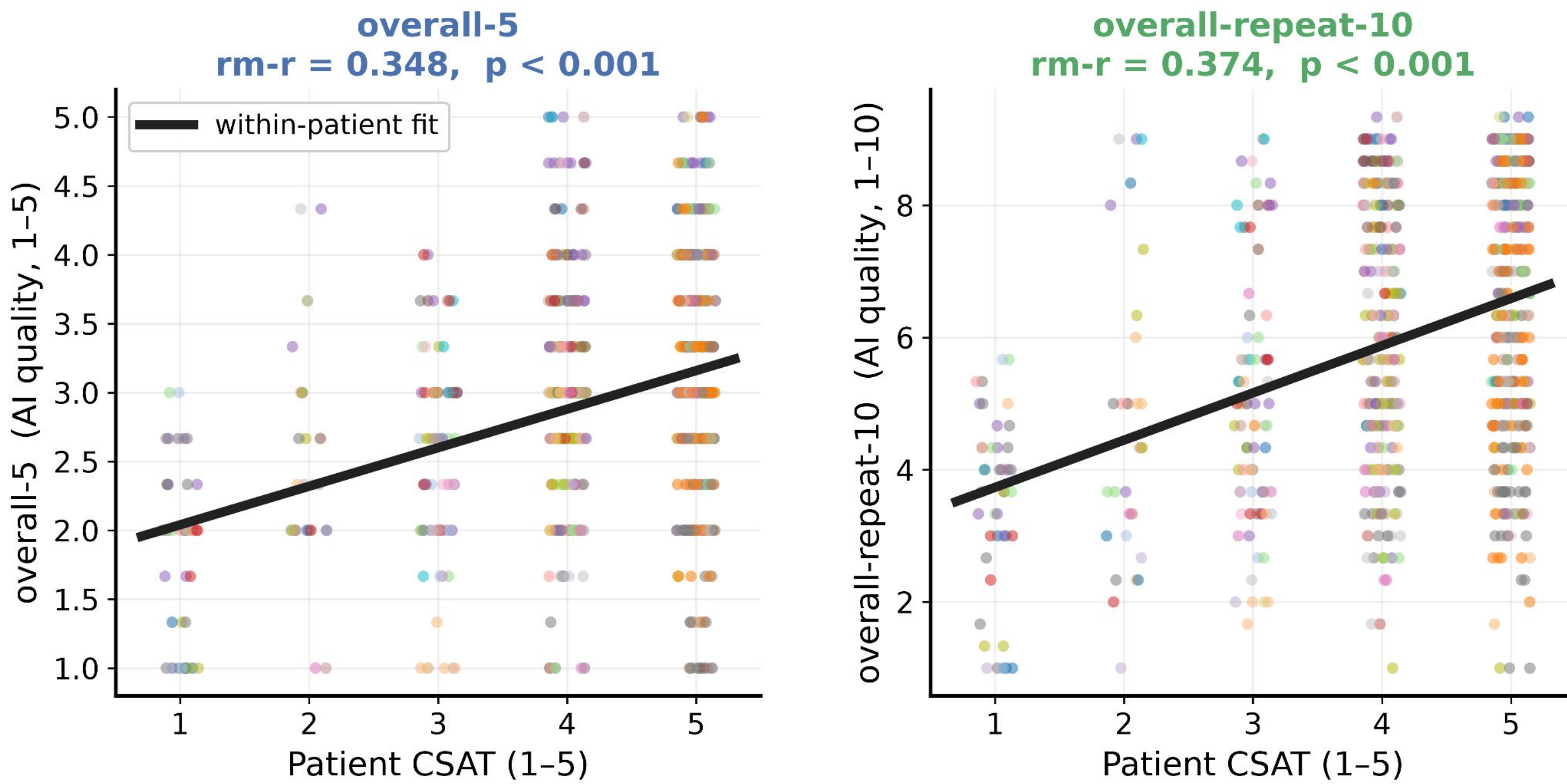


Figure 1. Within-patient agreement between automated score and patient CSAT (gpt-4o, t=0.7). Each dot = one call, colored by patient (36 patients, 729 calls); line = within-patient (repeated-measures) fit. Both evaluators track CSAT within patients: overall-5 rm-r = 0.348, overall-repeat-10 rm-r = 0.374 (both p < 0.001).

Results:

We compared each automated evaluator score against the patient's own post-call satisfaction rating (CSAT, 1–5) across 729 calls from 36 patients. Every one of the 17 judge configurations agreed significantly with patient satisfaction (all p < 0.001; Table 1), and critically, the agreement held within individual patients, not just across the group. Looking only at how a single patient's scores varied across their own calls, higher automated scores still matched higher CSAT (Figure 1), so the evaluator is tracking real call-to-call quality rather than just stable differences between patients. The best configuration was the overall-repeat-10 rubric scored by gpt-4o at temperature 0.7 (r = 0.383, within-patient r = 0.374), though three independently written rubrics converged on essentially the same answer. Notably, the choice of judge model mattered more than the rubric: gpt-4o outperformed gpt-4.1, claude-sonnet-4-6, and even the gpt-5.5 reasoning model, indicating that more reasoning did not make a better judge. A key check confirmed the signal is trustworthy: it was not an artifact of call length, as the correlation barely changed after controlling for transcript words and number of patient turns.

Discussion:

An LLM-as-judge can score conversational-AI call quality in a way that agrees with patients' own satisfaction, both across calls and within individual patients over time, making automated quality monitoring viable at scale. Notably, a mid-tier model (gpt-4o) outperformed a reasoning model, suggesting good judging depends more on calibration than on added reasoning. The agreement is moderate, as expected when validating against a single, positively skewed self-report item (79% of calls were rated 4 or 5 stars) that also reflects voice, speech-recognition, image-display and connectivity factors the transcript-only judge never sees. Both the comparison and the judge itself have room to improve. Next steps are to give the judge the signals it currently lacks (paralinguistic and technical-failure cues and a rebalanced rubric) and to validate against clinician-annotated quality rather than satisfaction alone.

References

1. Bahar-Fuchs et al. (2019). Cochrane Database Syst Rev.
2. Gavelin et al. (2020). Neuropsychol Rev.