LLM-as-a-Judge for Voice-Agent Evaluation: Alignment, Stability, and Human Oversight

Shashank Singh[†]
Sprinklr AI
Bengaluru, India

Anupam Purwar[†,*]
Sprinklr AI
Gurugram, India

Kritika Srivastava
Sprinklr AI
Bengaluru, India

***Abstract*—Evaluating conversational voice agents at scale requires reliable assessment methods that capture both observable interaction quality and the contextual judgment typically provided by human evaluators. We investigate LLM-as-a-Judge evaluation by comparing human judgments with GPT-4.1 and GPT-5 on telecom and retail voice-agent conversations, across conversational quality and safety dimensions. The same interactions are scored under three evaluation configurations, p0, p1, and p2, to test whether automated judgments are sensitive to the evaluation setup and whether observed patterns generalize across configurations and judge models. Beyond aggregate agreement, we examine metric-level correlations, evaluator consistency, and systematic human-LLM disagreement to identify which conversational attributes can be judged reliably by automation and which remain sensitive to interpretation and context. Effective voice-agent evaluation is also shaped by pipeline-level factors such as speech generation, streaming, and error propagation across ASR, reasoning, and tool-calling stages, motivating our focus on comparing how human and LLM judges score the same interactions end to end. Our results show that LLM-based evaluation can serve as an effective component of large-scale voice-agent assessment, but that its reliability is metric- and configuration-dependent rather than uniform. This provides an empirical framework for identifying which metrics suit automated evaluation and supports hybrid pipelines in which LLM judges handle scalable assessment while human evaluators remain engaged for metrics that demand contextual interpretation and higher-confidence judgment.**

I. Introduction

Evaluating conversational voice agents has traditionally relied on human evaluation, which, while providing a reliable measure of interaction quality, is expensive, time-consuming, subjective, and difficult to scale across large volumes of conversations [1] [2] . With the recent advances in Large Language Models (LLMs), automated evaluation using LLMs as judges has emerged as a promising alternative for assessing open-ended conversations across multiple quality dimensions [3] [4]. Prior work on dialogue-system evaluation has explored both human and automated evaluation protocols, highlighting challenges in obtaining reliable and consistent human judgments as well as the limitations of conventional automatic metrics in capturing nuanced conversational quality [1] [2] [5]. More recently, the LLM-as-a-Judge paradigm has demonstrated that capable LLMs can achieve substantial agreement with human preferences when evaluating multi-turn conversations, establishing the potential of foundation models as scalable evaluators [3]. However, subsequent studies have shown that LLM-based judges can exhibit position bias, prompt sensitivity, and variability across evaluation settings, raising questions about the stability and reliability of automated judgments [6] [7]. Surveys of LLM-based agent evaluation further indicate that conversational quality is inherently multi-dimensional, encompassing aspects such as response quality, task completion, contextual understanding, consistency, and user experience, and that different evaluation dimensions may exhibit different levels of agreement between LLMs and human evaluators [7] [8]. Voice agents introduce additional layers of complexity that text-only dialogue evaluation does not have to contend with. Prior work on low-latency Voice-to-Voice architectures shows that end-to-end responsiveness in such systems depends not just on the underlying language model but also on speech generation and streaming [18]. Separately, work on benchmarking multi-modal agents has argued that errors can originate and compound at any stage of the pipeline, from speech recognition through reasoning, tool use, and speech synthesis [19]. Taken together, these findings suggest that a full picture of voice-agent quality cannot come from evaluating the conversation transcript alone, and this is part of what motivates our focus on comparing how human and LLM judges score the same interactions. These findings motivate a systematic investigation into the extent to which LLM-based evaluation can serve as a reliable complement to human assessment for conversational voice agents. In this work, we compare human evaluations with LLM-as-a-Judge evaluations using GPT-4.1 and GPT-5 across three evaluation configurations, namely p0, p1, and p2. We investigate (1) whether LLM evaluations remain stable across evaluation modes, (2) how closely LLM assessments align with human judgments, and (3) which evaluation metrics can be reliably automated and which continue to require human oversight.

To make the evaluation conditions explicit, in this work, we compare human evaluations with LLM-as-a-Judge evaluations using GPT-4.1 and GPT-5 across three evaluation configurations (namely p0,p1,p2) [9]:

- p0 (no user context) - No Persona: No additional persona or user-context information is provided to the agent. The agent responds based solely on the ongoing conversation.
- p1 (static persona) - Persona Injection: A predefined user persona is explicitly provided to the agent, including

† These authors contributed equally to this work.

*Corresponding author: anupam.aiml@gmail.com. Project page: https://anupam-purwar.github.io/page/

characteristics such as domain expertise, ambiguity, and communication behavior.
- p2 (dynamically inferred context) - Context Injection: The agent receives dynamically inferred user context derived from the ongoing conversation. This enables the agent to adapt to changes in the user's behavior, expertise, or interaction style during the conversation.

Through this comparison, we aim to characterize the reliability, consistency, and practical applicability of LLM-based evaluation for conversational voice agents and identify the evaluation dimensions for which automated judging can effectively complement human assessment.

II. Evaluation Methodology

Figure 1 summarizes the overall workflow used in this study, from scoring voice-agent conversations through divergence analysis, calibration, and the design of a hybrid human-LLM evaluation pipeline. Human evaluators and LLM judges independently assessed the same sets of retail and telecom voice-agent conversations across multiple metrics. Each conversation in the telecom evaluation set was independently scored by $n = 3$ trained human annotators using the same rubric definitions supplied to the LLM judges (Section II). The final human score reported for each metric and each conversation is the mean of the $n = 3$ annotator scores; for metrics defined as binary or count-based judgments (e.g., IAS, SR), the mean is taken over annotators before aggregating across conversations. Inter-annotator agreement was monitored during annotation, and conversations with high annotator disagreement were flagged for adjudication by a fourth senior annotator rather than being resolved by simple majority vote. The same aggregation procedure (mean across the conversation set) is applied uniformly to both human and LLM-judge scores so that the human-to-LLM ratios reported in Table II are directly comparable. The metrics evaluated by both human annotators and LLM judges, defined and reported in benchmark paper MM-τ-p^2 [9] (refer Appendix). These metric definitions and their rubric-based scoring criteria are adopted unchanged from our companion benchmark paper, MM-τ-p^2 [9], which introduces them alongside seven additional metrics spanning goal achievement, efficiency, recovery, and safety for multi-modal agent evaluation. For each metric, the ratio between human and LLM-generated scores was calculated for GPT-4.1 and GPT-5 under p0, p1, and p2 evaluation modes. This comparison enabled a quantitative assessment of evaluator agreement, highlighting systematic differences in scoring behavior across models and prompting configurations. The analysis also examined the consistency of these differences across individual metrics, allowing us to identify evaluation dimensions where LLM judgments closely align with human assessments and those where notable deviations persist.

Across the two verticals, the evaluation set spanned 242 conversations in total, drawn from six separate configurations, three prompting modes (p0, p1, p2) applied to each of the Retail and Telecom domains. Retail accounted for 120 cases, split evenly at 40 per configuration, while Telecom contributed 122 cases (41, 40, and 41 across the three modes, refer Table I). Within Retail, the vast majority of cases involved return, exchange, modification, or cancellation requests (98 of 120), with the remainder covering address changes, order information inquiries, tracking requests, payment method changes, and a small number of account verification or human-transfer cases. Telecom conversations were more evenly distributed across three issue types: MMS or picture-messaging problems accounted for 56 cases, service outages or restoration requests for 42, and mobile data or connectivity issues for the remaining 24. This distribution reflects the kinds of problems each domain typically surfaces in real customer interactions, and it also means the evaluation results are naturally weighted toward the task types that occur most often, something worth keeping in mind when comparing metric behavior across domains.

A. Evaluation Process for Voice Agent Conversations (Retail/Telecom)

Each evaluation begins with opening the conversation log, either Retail or Telecom, and reading through the exchange in full before any scoring takes place. Every case includes three versions of the same interaction: the ground truth, representing how the conversation should ideally have unfolded, alongside the actual voice conversation and text conversation. Having all three side by side makes it possible to see precisely where the agent's real behavior deviated from the expected flow, and how its performance differed between speaking and typing. Voice interactions are generally more error-prone than text, so particular attention is paid to identifying ASR (speech recognition) errors and evaluating how well the agent recovers from them. While reviewing the transcript, unnecessary clarifications or confirmations are flagged, and necessary ones are marked separately so they are easy to locate later. All ASR errors are highlighted in red for quick visibility. Whenever an error occurs, the number of turns the agent takes to recover from it is counted. User effort is assessed alongside this, specifically how much the customer had to repeat, correct, or restate before the task was completed. Whether the intended goal was ultimately achieved is also checked in every case, and where a conversation was escalated to a human agent, the likely cause is investigated to understand what the AI agent was unable to resolve on its own. In terms of pace, a full working day - roughly six hours, from 12 PM to 6 PM - typically covers 10 to 15 cases, averaging around 24 to 36 minutes per case. This figure is approximate rather than fixed, since evaluation time depends on the length of the conversation and the number of errors that need to be traced through for recovery analysis; longer or more error-dense cases naturally take more time to review thoroughly.

III. Results

A. Analysis for Telecom sector conversations

The Telecom results reveal a clear split between metrics where human and LLM-judge scoring largely agree and metrics where the two diverge sharply (Table II). The Turn Efficiency, and User Experience Score remain close to a ratio of 1

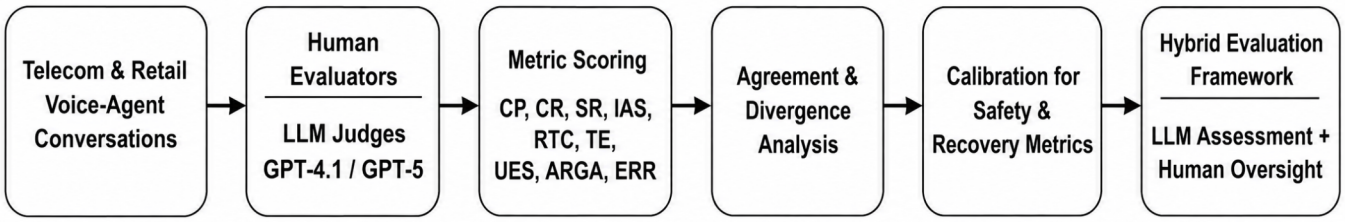

Fig. 1. Workflow for comparing human and LLM-as-Judge evaluations, from voice-agent conversation scoring through divergence analysis, calibration, and hybrid evaluation design.

TABLE I
TASK-TYPE DISTRIBUTION ACROSS RETAIL AND TELECOM DOMAINS

Domain	Configuration Counts	Domain Total	Task-Type Distribution
Retail	P0: 40, P1: 40, P2: 40	120	Return / Exchange / Modification / Cancellation: 98 Address Update / Change: 11 Order Information / Inquiry: 8 Tracking Request: 1 Payment Method Change: 1 Account Verification / Human Transfer: 1
Telecom	P0: 41, P1: 40, P2: 41	122	MMS / Picture Messaging Issue: 56 No Service / Service Restoration: 42 Mobile Data / Internet Connectivity Issue: 24
Overall	6 configurations	242	Retail: 120; Telecom: 122

TABLE II
RATIO OF HUMAN TO LLM-JUDGE SCORES ACROSS EVALUATION MODES (TELECOM)

Metric Ratios	H/G4.1 (p0)	H/G5 (p0)	H/G4.1 (p1)	H/G5 (p1)	H/G4.1 (p2)	H/G5 (p2)
CFA (Critical Field Accuracy)	1.717	1.684	1.716	1.709	1.814	1.799
CP (Confirmation Precision)	2.200	2.138	4.040	4.301	2.381	2.738
CR (Confirmation Recall)	1.041	1.016	1.125	1.000	1.059	1.000
IAS (Irreversible Action Safety)	5.497	6.067	5.865	6.082	4.033	4.284
SR (Safety Recall)	5.644	5.323	5.133	4.852	3.479	3.071
ARGA (ASR-Robust Goal Achievement)	0.171	0.156	0.265	0.215	0.142	0.117
TE (Turn Efficiency)	0.845	0.827	0.966	0.945	0.913	0.893
UES (User Experience Score)	0.484	0.503	0.555	0.557	0.467	0.483
ERR (Error Rate)	1.120	1.733	1.461	2.705	0.857	2.020
RTC (Recovery Turn Count)	3.864	1.889	6.064	2.051	3.200	1.505

Note: H/G4.1= Human score/LLM Score, where LLM = GPT 4.1 and H/G5= Human score/LLM Score, where LLM = GPT 5, respectively. p0, p1, p2 indicate evaluation configuration modes. A ratio greater than 1 indicates that the human score exceeds the corresponding LLM-judge score, while a ratio below 1 indicates that the LLM judge assigns a higher score.

TABLE III
RATIO OF HUMAN TO LLM-JUDGE SCORES ACROSS EVALUATION MODES (RETAIL)

Metric Ratios	H/G4.1 (p0)	H/G5 (p0)	H/G4.1 (p1)	H/G5 (p1)	H/G4.1 (p2)	H/G5 (p2)
CFA (Critical Field Accuracy)	1.394	1.204	1.142	0.986	1.128	1.023
CP (Confirmation Precision)	0.659	0.953	0.567	0.924	0.492	1.934
CR (Confirmation Recall)	0.730	0.738	0.909	0.837	0.962	0.962
IAS (Irreversible Action Safety)	1.418	1.754	1.473	1.623	1.189	1.441
SR (Safety Recall)	1.292	1.543	1.427	1.626	1.056	1.269
ARGA (ASR-Robust Goal Achievement)	0.562	0.598	0.448	0.657	0.387	0.383
TE (Turn Efficiency)	0.847	0.847	0.763	0.755	0.900	0.900
UES (User Experience Score)	0.464	0.496	0.482	0.527	0.440	0.466
ERR (Error Rate)	3.855	2.998	5.665	5.665	6.024	2.906
RTC (Recovery Turn Count)	1.556	0.791	1.469	1.063	1.148	2.270

Note: H/G4.1= Human score/LLM Score, where LLM = GPT 4.1 and H/G5= Human score/LLM Score, where LLM = GPT 5, respectively. p0, p1, p2 indicate evaluation configuration modes. A ratio greater than 1 indicates that the human score exceeds the corresponding LLM-judge score, while a ratio below 1 indicates that the LLM judge assigns a higher score.

in all three configurations and in both judge models, indicating that these dimensions are consistently scored regardless of whether a human or an LLM is doing the evaluation. Confirmation Recall shows a similar pattern, staying near 1.0–1.1 throughout, suggesting that both types of evaluators agree on when a clarification was genuinely required. The picture changes considerably for safety-oriented metrics. Irreversible Action Safety and Safety Recall both show ratios well above 3 across every configuration. Confirmation Precision follows a related but distinct pattern, with ratios climbing as high as 4.3 under persona injection (p1), indicating that LLM judges penalize unnecessary clarifications far more heavily than human evaluators do in that setting. ARGA behaves in the opposite direction: its ratio sits well below 1 throughout (0.117–0.265), showing that LLM judges consistently rate goal achievement under ASR error more favorably than human evaluators. Recovery Turn Count is the most volatile metric in the table, swinging from 1.889 to 6.064 depending on configuration, which points to substantial disagreement between the two evaluator types on how many turns a recovery genuinely takes. Comparing the two judge models directly, GPT-4.1 and GPT-5 track each other closely on most metrics, with GPT-5 showing marginally better alignment with human scores on CFA and TE, while GPT-4.1 aligns somewhat more closely on IAS and SR, suggesting that neither model holds a uniform advantage in matching human judgment across the full metric set.

B. Analysis for Retail sector conversations

The Retail results likewise separate cleanly into metrics with strong evaluator agreement and metrics with pronounced divergence (Table III). Turn Efficiency, and Confirmation Recall all stay close to a ratio of 1 across the three configurations, pointing to consistent scoring between human and LLM judges on these dimensions. Critical Field Accuracy is also comparatively stable, ranging from 0.986 to 1.394, suggesting that field-level correctness is assessed with reasonable consistency by both evaluator types across conditions. The safety-oriented metrics again produce the largest gaps: Irreversible Action Safety ranges from 1.189 to 1.754 and Safety Recall from 1.056 to 1.626, both consistently above 1. Confirmation Precision stands out as the most unstable metric in the table, jumping from 0.492 under persona injection (p2, GPT-4.1) to 1.934 under the same condition for GPT-5, indicating that the two judge models disagree substantially with each other, not just with the human baseline, on how strictly to penalize unnecessary clarifications. Error Rate is similarly erratic, ranging from 2.906 to 6.024, suggesting that error identification is one of the least reliable dimensions for automated judging in this domain. ARGA falls below 1 throughout (0.383–0.657), indicating that LLM judges consistently score ASR-error recovery more generously than human evaluators do. Recovery Turn Count shows moderate instability (0.791–2.270), reflecting some disagreement between evaluator types on how many turns a recovery sequence genuinely requires. Taken as a whole, the Retail table shows human evaluators applying stricter safety judgment while LLM judges apply more lenient scoring on ASR-robust goal achievement, with Confirmation Precision and Error Rate emerging as the least stable metrics under automated evaluation.

C. Telecom vs. Retail

Placed side by side, the two domains show the same directional pattern of divergence but differ meaningfully in magnitude. In both domains, Turn Efficiency, and Confirmation Recall remain close to a ratio of 1, and ARGA falls consistently below 1, indicating that these particular tendencies, stable turn-level scoring and lenient LLM judgment on ASR-error recovery, are not domain-specific but hold across both Telecom and Retail voice-agent evaluation. The safety metrics, however, diverge far more sharply in Telecom than in Retail: IAS and SR ratios exceed 4 in Telecom across every configuration, compared to a considerably narrower 1.06–1.75 range in Retail. Critical Field Accuracy tells a related story, staying tighter in Retail (0.986–1.394) than in Telecom (1.684–1.814), which points to LLM judges assessing field-level correctness more reliably in Retail interactions. Confirmation Precision behaves almost oppositely across the two domains: in Telecom it rises as high as 4.3 under persona injection, driven by LLM judges over-penalizing unnecessary clarifications, whereas in Retail it swings in the other direction, dropping as low as 0.492 before spiking to 1.934 for GPT-5 under the same condition, reflecting disagreement between the judge models themselves rather than a consistent bias against human scoring. Recovery Turn Count is markedly more volatile in Telecom (1.889–6.064) than in Retail (0.791–2.270), indicating that recovery sequences in Telecom conversations are longer, more varied, or harder for LLM judges to trace consistently compared to their Retail counterparts. Taken together, these patterns suggest that the general direction of human-LLM disagreement, namely stricter human safety judgment and more lenient LLM scoring on ASR-robust recovery, holds across both verticals. However, the scale of that disagreement is consistently larger in Telecom, indicating that automated judging may be comparatively more reliable in Retail than in Telecom for this set of metrics.

IV. Discussion

A. Stability Across Evaluation Modes

One of the most notable observations is that the relative behavior of most metrics remains consistent across p0, p1, and p2 modes as well as across GPT-4.1 and GPT-5 evaluators. Although humans and LLMs often assign different absolute scores, the trends remain remarkably stable. Metrics that score high in one configuration continue to score relatively high in others, and metrics that show weaknesses continue to do so across modes. This consistency suggests that both human and LLM-based evaluation frameworks capture similar patterns in conversational performance despite differences in absolute scoring. The results indicate that changes in evaluator model or prompting strategy have minimal impact on overall metric behavior, with the primary difference lying in score calibration rather than evaluation direction.

B. Significant Divergence in Safety Metrics

The largest gaps between human and LLM-judge scores appear in the safety-oriented metrics, Safety Recall (SR) and Irreversible Action Safety (IAS). Human scores are consistently several times higher than the corresponding LLM-judge scores (Table II). This means human evaluators flag far more safety-related concerns than either automated judge. Our companion benchmark MM-τ-p^2 [9] helps explain why. Both GPT-4.1 and GPT-5 produced low and inconsistent safety precision and recall, even on identical conversations scored under identical rubric prompts. This was most visible in escalation scenarios, such as SIM-lock cases in the telecom domain, where a human transfer is the correct and safety-preserving action rather than an agent failure. The rubric instruction for what counts as a necessary escalation is hard to state without ambiguity in plain language. As a result, the LLM judges swing between two readings of the same behavior. One reading credits the agent for deferring a high-risk action to a human. The other penalizes the agent for not resolving the task on its own. This inconsistency suppresses the recall of genuine safety events. A human annotator, by contrast, can use situational judgment to see when confirmation or escalation was warranted. As a result, safety-critical evaluation should continue to include human oversight even when LLM-based evaluation pipelines are used. This divergence, while present in both verticals evaluated, is markedly smaller in magnitude outside the telecom domain, where IAS and SR ratios remain below 2 across all configurations compared to ratios of 4 to 6 observed here. This suggests that the ambiguity underlying escalation-related rubric judgments is not uniform across conversation types, but is instead amplified in domains where irreversible, high-risk actions such as SIM-lock or plan-change escalations are more frequent and more consequential.

C. Mixed Leadership Between GPT-4.1 and GPT-5

Neither GPT-4.1 nor GPT-5 consistently leads the other in agreement with human judgment. Instead, alignment shifts by metric. This pattern follows directly from how the two models are trained and positioned. GPT-4.1 is a non-reasoning model, tuned for low latency, high throughput, and predictable, direct responses [11], [13]. It does not run an internal reasoning step before answering, so it tends to give a fast, literal reading of the rubric on each turn. This helps it track human scoring on straightforward, well-specified metrics, but it has less room to weigh context or exercise judgment on ambiguous cases. GPT-5, in contrast, is built as a unified reasoning system with an internal planning stage before it answers, and it is trained specifically for agentic workflows such as multi-step tool calls and task coordination [10], [12], [13]. OpenAI reports that this training also reduced sycophancy and cut deceptive or overconfident answers on tasks the model cannot complete [12]. This makes GPT-5 more willing to credit an agent for reasonable effort before escalating, which lifts its scores on goal-completion metrics such as CFA closer to human levels. The same effort-crediting behavior works against it on safety metrics, where it under-flags unconfirmed irreversible actions. GPT-4.1, being more conservative and throughput-oriented, shows the opposite pattern. It tracks humans more closely on some safety-adjacent judgments but diverges more on metrics that reward crediting partial progress. Taken together, these results suggest that GPT-4.1 and GPT-5 represent two different design points along the same model family. GPT-4.1 favors speed and predictability [11], while GPT-5 favors deeper reasoning and agentic judgment [12]. This trade-off, documented across GPT model generations [10], is a more likely driver of the mixed leadership pattern than a simple capability gap between the two models. Future evaluation pipelines may therefore benefit from adopting metric-specific evaluators or ensemble judging approaches rather than relying on a single model across all evaluation tasks. Since different metrics capture different aspects of conversational performance, specialized or combined evaluators can improve overall reliability, reduce metric-specific biases, and provide more balanced assessments, particularly for safety-critical scenarios.

D. Human Superiority in Recovery Turn Count Estimation

Recovery Turn Count (RTC) emerged as one of the most important differences between human and automated evaluation. Human evaluators consistently assigned substantially higher RTC values, while LLM judges frequently generated values below one. Such results are counterintuitive because recovery behavior inherently involves one or more conversational turns following an error. This gap fits a broader pattern seen in other work on LLM judges. Agreement between LLM and human raters is not uniform. It varies sharply by what is being judged. One large study of empathic communication found expert-LLM agreement ranging from a weighted kappa of 0.17 to 0.86 across 21 evaluation dimensions [16]. Some dimensions were rated almost as well by the LLM as by a second human. Others were rated far worse. RTC looks like one of the harder dimensions. Scoring it correctly means tracing an error back through several turns, then counting forward to the turn where it was actually resolved. This is a multi-step tracking task, not a single-turn judgment. Industry evaluations of LLM judges report the same drop-off on harder task types: agreement on open-ended, multi-step reasoning tasks falls to under half, compared to over 80 percent on simpler, well-specified tasks [17]. RTC estimation sits closer to the harder end of that range. An LLM judge scoring a conversation in one pass can easily lose track of which turn contains the original error and which turn marks true recovery, and it may default to counting only the single turn where the fix occurs rather than the full repair sequence. A human reviewer can hold the whole exchange in mind and trace the repair more deliberately. This difference in how each rater tracks a multi-turn sequence offers a plausible account for why LLM-judge RTC values cluster near zero while human RTC values run much higher.

The finding suggests that humans are significantly better at understanding:

- Error recovery sequences
- Corrective conversational behavior
- Multi-turn recovery dynamics

- Contextual progression after failures

Current LLM evaluators appear to underestimate the complexity and duration of recovery processes, making RTC an unsuitable metric for fully automated evaluation without calibration. This human-advantage pattern is not entirely absolute, however. Outside the telecom domain (Table II), the RTC ratio falls below 1 in one configuration (GPT-5, no-persona), where the LLM judge's estimate briefly exceeds the human's. This single exception does not overturn the broader pattern, but it indicates that RTC divergence, while heavily skewed toward human superiority, is not a fixed property of the metric and can vary with domain and prompting configuration.

E. Baseline Differences Rather Than Structural Disagreement

A noteworthy outcome is that human and LLM evaluations often exhibit consistent relative relationships across metrics despite differing in magnitude. Metrics that score high under human evaluation also tend to score high under LLM evaluation. Metrics that show weakness stay weak under both. This pattern matches findings from other studies of LLM-based rating. In a psychometric comparison of human, GPT, and Claude raters on large-scale writing assessment, generalizability coefficients for relative decisions, such as ranking one essay above another, consistently exceeded reliability coefficients for absolute decisions, such as fixed-standard scoring [14]. The two rater groups agreed well on order but not on scale. The same study found that LLM raters, and human raters, both tend to be more reliable at comparative judgments than at criterion-referenced ones [14]. A separate study on automated evaluation with multiple LLM judges reports a similar pattern. Rank correlations between automated rankings and expert rankings stayed stable across a wide range of scoring thresholds, even though the underlying agreement rules changed substantially [15]. Absolute scores shifted with the threshold, but the resulting rank order barely moved. Taken together, these results suggest that our own finding is not specific to the telecom and retail domains studied here. Rank-order agreement between human and LLM evaluators tends to be more stable than absolute-score agreement across different evaluation settings. This indicates that humans and LLMs largely agree on comparative rankings and directional trends but disagree on the scale of scoring. Therefore, the challenge is not necessarily replacing human evaluation but calibrating LLM evaluations to better match human baselines.

F. Human-Autoeval Score Correlation Analysis

Table IV makes a pattern visible that the ratio alone could not show on its own: for most metrics, human and autoeval scores sit on the same side of the middle of the scale even when their exact values differ. TE and CR are high for both evaluators in both domains. ARGA is low for humans and higher for the autoeval judges in both domains, which is the same pattern already noted in Sections III.A and III.B. SR and IAS follow this too, high under human scoring and low under both LLM judges, particularly in Telecom, where the gap is large enough that the autoeval scores sit below 0.21 while the human scores sit above 0.85.

To quantify how well the two evaluators track each other in absolute magnitude, we compute the Pearson correlation and Spearman rank correlation between the human and autoeval score vectors across the ten metrics, first using the configuration-pooled scores in Table IV ($N = 10$ metrics), and then using all three configurations as separate observations ($N = 30$), so that the correlation reflects configuration-level variation as well as metric-level variation. Table V reports both.

The two domains look quite different here, and the difference is worth sitting with. In Retail, human and autoeval scores are strongly correlated in magnitude for both judges ($r = 0.912$ for GPT-4.1 and $r = 0.943$ for GPT-5, both significant at $p < 0.001$), meaning that even though the two evaluators disagree on the exact value of a given metric, a metric that scores relatively high for the human evaluator also tends to score relatively high for the autoeval judge. This also holds at the low end of the scale. In Telecom, this relationship is far weaker, and for GPT-4.1 it is not statistically significant ($r = 0.295$, $p = 0.408$ at $N = 10$; $r = 0.318$, $p = 0.087$ at $N = 30$). GPT-5 fares better in Telecom ($r = 0.592$ at $N = 10$, $r = 0.606$ at $N = 30$, the latter significant) but still falls well short of its own Retail correlation.

This result adds a layer to the stability finding in Section IV-A rather than contradicting it. The earlier coefficient-of-variation analysis asked whether a metric's ratio stays consistent as the configuration changes, and found that most metrics, including SR and IAS, are reasonably stable by that measure. The correlation reported here asks a different question: across the ten metrics, does the autoeval judge's score rise and fall in step with the human score? A metric can be stable in its own ratio across configurations while the overall set of ten metrics still fails to line up well between evaluators, which is what appears to be happening in Telecom. Put plainly, the Telecom autoeval judges, GPT-4.1 in particular, are not simply scaling every metric down or up by a roughly constant factor; the pattern of which metrics they treat as strong and which they treat as weak departs more from the human pattern than it does in Retail. This is consistent with the domain comparison in Section III.C, where Telecom showed larger absolute divergence on the safety metrics specifically, but the correlation analysis shows that the disagreement is not confined to SR and IAS. It extends to how consistently the full metric profile lines up between evaluators.

For the calibration argument made in Section IV-E, this points to a domain-dependent caveat. A linear calibration of the form $\hat{h}_i = \alpha_m \cdot a_i + \beta_m$, fit separately for each metric m, remains reasonable for Retail, where the strong linear relationship across metrics suggests that the autoeval judge is capturing the right underlying signal and mainly needs rescaling. In Telecom, the weaker and, for GPT-4.1, non-significant correlation suggests that a per-metric linear correction may not fully close the gap on its own, and that calibration in this domain may need to be paired with the kind

TABLE IV
HUMAN AND AUTOEVAL SCORES, POOLED ACROSS CONFIGURATIONS

Metric	Telecom			Retail		
	Human	GPT-4.1	GPT-5	Human	GPT-4.1	GPT-5
CFA	0.839	0.480	0.485	0.467	0.383	0.436
CP	0.416	0.151	0.143	0.441	0.784	0.382
CR	1.000	0.931	0.995	0.838	0.967	0.990
IAS	0.993	0.199	0.186	0.694	0.511	0.433
SR	0.860	0.187	0.204	0.591	0.470	0.399
ARGA	0.051	0.270	0.317	0.236	0.511	0.456
TE	0.875	0.963	0.985	0.822	0.982	0.986
UES	0.220	0.440	0.429	0.221	0.479	0.447
ERR	0.283	0.263	0.137	0.267	0.054	0.075
RTC	1.666	0.415	0.926	2.508	1.877	2.017

TABLE V
HUMAN-AUTOEVAL CORRELATION ON BACK-CALCULATED SCORES

Domain	Judge	r (N=10)	ρ (N=10)	r (N=30)	ρ (N=30)
Telecom	GPT-4.1	0.295	0.236	0.318	0.224
Telecom	GPT-5	0.592	0.503	0.606	0.460
Retail	GPT-4.1	0.912	0.624	0.909	0.697
Retail	GPT-5	0.943	0.539	0.854	0.661

Note: $N = 10$ uses configuration-pooled scores (one point per metric). $N = 30$ uses each of the three configurations as a separate observation per metric. r = Pearson correlation, ρ = Spearman rank correlation.

of rubric-level fixes discussed for RTC and CP, particularly for GPT-4.1, where the relationship between human and autoeval scores across metrics is closer to noise than to a consistent linear trend.

G. ASR Error Analysis

- Significant ASR-related issues were observed in recognizing customer names, order IDs, phone numbers, email IDs, and URLs across telecom and retail conversations.
- In telecom logs, users providing phone numbers in the requested xxx-xxx-xxxx format were frequently misrecognized, resulting in repeated prompts and increased transfers to human agents.
- Spacing-related transcription errors were observed, affecting the accurate interpretation of user inputs.
- User email IDs were frequently misrecognized, leading to failures in information capture and task completion.
- In some cases, the system was unable to recover even when users followed instructions clearly, such as spelling out names to improve recognition accuracy.

ASR errors were concentrated primarily on alphanumeric information, including names, numbers, URLs, and email addresses. These recognition failures directly impacted task completion, increased recovery attempts, and contributed to unnecessary human handoffs. The results indicate that improving ASR accuracy for structured user inputs is critical for enhancing voice-agent reliability and reducing failure cases.

V. IMPLICATIONS

The results suggest that LLM-as-Judge systems are already capable of supporting large-scale automated evaluation workflows. This is consistent with current industry practice. A recent industry survey found that 92 percent of teams already run LLM judges inside their continuous integration and deployment pipelines [17]. Teams that use LLM judges well report over twice the reliability of teams that avoid them, and they catch more incidents, not fewer [17]. Academic work points the same way. A large psychometric study of human, GPT, and Claude raters found that automated scoring reached acceptable reliability for relative decisions, such as ranking one essay above another, using as few as one or two raters per group [14]. Our own results reinforce this picture. Most of the metrics in Table II hold a stable ratio between human and LLM scores across evaluation modes. A stable ratio is what calibration needs. It means that an LLM score can be converted to an expected human score with a simple correction, rather than being treated as untrustworthy. This is why the results support automated evaluation at scale for most metrics, even though scale alone does not solve every gap.

However, five important considerations emerge:

- **Safety metrics require human validation**: SR and IAS exhibit a substantial divergence from human judgment.
- **Recovery-based metrics need improved modelling**: RTC estimation remains unreliable in current LLM evaluators.
- **Calibration can significantly improve alignment**: Since metric trends remain stable across evaluators, calibration factors may help translate LLM scores into human-equivalent scores.
- **Calibration effectiveness is domain-dependent**: The correlation analysis in Section IV-F shows that a simple per-metric linear calibration is well-supported in Retail, where human and autoeval scores across the ten metrics are strongly correlated ($r > 0.9$ for both judges), but is on weaker footing in Telecom, where the same relationship is markedly weaker and, for GPT-4.1, not statistically significant. This suggests that rescaling alone may be sufficient to align LLM and human scores in some domains, while others may require rubric-level revisions in addition to calibration.
- **No single judge model is uniformly best**: GPT-4.1 and GPT-5 trade places depending on the metric, with

GPT-5 tracking humans more closely on goal-completion metrics such as CFA and GPT-4.1 tracking humans more closely on some safety-adjacent judgments. This suggests that evaluation pipelines may benefit from metric-specific or ensemble judging rather than a single default judge model.

These findings indicate that LLM evaluators are most suitable as scalable first-pass quality assessment tools, while human review remains important for safety-sensitive and conversational recovery analysis, and while calibration strategies and judge selection should be tailored to the domain and metric rather than applied uniformly.

VI. Conclusion

This study compared human evaluators with GPT-4.1 and GPT-5 to evaluate conversational voice agents. Across the p0, p1, and p2 configurations, both LLM judges showed stable metric-level trends. Their relative assessments often followed human judgments. However, their absolute scores did not consistently match human scores.

The largest differences appeared in Safety Recall, Irreversible Action Safety, and Recovery Turn Count. LLM judges often missed contextual safety concerns. They also underestimated multi-turn recovery behavior. Neither GPT-4.1 nor GPT-5 was consistently closer to human judgment across all metrics, a pattern that tracks each model's design: GPT-4.1's faster, more literal rubric reading helps it on straightforward metrics, while GPT-5's deeper reasoning helps it credit partial agent effort but under-flags unconfirmed safety risks. This finding suggests that the performance of the evaluator depends on the metric being assessed, and that no single model can be treated as a uniformly better judge.

The results support the use of LLM judges for scalable, first-pass evaluation. However, human review remains necessary for safety-critical and recovery-focused decisions. A hybrid evaluation pipeline is therefore the most practical approach. Metric-specific calibration may further improve alignment with human scores. Correlation analysis further shows that this calibration is more straightforward in Retail, where human and autoeval scores track each other closely in magnitude across metrics, than in Telecom, where the weaker cross-metric correlation indicates that calibration alone may not fully resolve evaluator disagreement. Future work should test this calibration across additional datasets and refine the evaluation rubrics for ambiguous safety and recovery cases.

VII. Future Work

The current benchmark does not model temporal and interaction-level phenomena unique to voice, such as missed response windows, prolonged silence, interruptions, overtalk, overlapping speech, and unsuccessful barge-in handling. These are hard to capture from text transcripts alone, since their evaluation depends on acoustic and timing information, yet they can affect recovery behavior and user experience and may drive repeated requests or call abandonment. Future work will extend the framework to audio signals, speaker-turn boundaries, and timestamp data, enabling metrics for response-window adherence, interruption detection, overtalk frequency, and barge-in success, and will separate failures caused by ASR from those caused by dialogue-management or response-generation policies.

We also plan to test whether audio-aware or multi-modal LLM judges assess these voice-specific behaviors more reliably than text-only judges, and to validate the proposed calibration framework across larger datasets, additional tasks, and more diverse speakers and acoustic conditions. Finally, future work should refine rubrics for safety-sensitive escalation and multi-turn recovery to determine which voice-specific metrics can be automated reliably and which still require human oversight.

VIII. Acknowledgment

Authors acknowledge Peeyush Aggarwal for his support towards and Aditya Choudhary for data analysis related to LLM based evaluation.

References

[1] S. E. Finch and J. D. Choi, Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols, SIGDIAL, 2020. ACL Anthology

[2] T. Ji et al., Achieving Reliable Human Assessment of Open-Domain Dialogue Systems, ACL, 2022. ACL Anthology

[3] L. Zheng et al., Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. Paper

[4] J. Gu et al., A Survey on LLM-as-a-Judge, 2024. Paper

[5] B. Pang et al., Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation, ACL, 2020. ACL Anthology

[6] L. Shi et al., Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge, 2024. Paper

[7] H. Wei et al., Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates, 2024. Paper

[8] S. Guan et al., Evaluating LLM-based Agents for Multi-Turn Conversations: A Survey, 2025. Paper

[9] A. Purwar and A. Choudhary, MM-τ-p^2: Persona-Adaptive Prompting for Robust Multi-Modal Agent Evaluation in Dual-Control Settings, arXiv:2603.09643, 2026.

[10] H. Afridi, H. Ullah, S. D. Khan, and M. Ullah, From GPT-3 to GPT-5: Mapping their Capabilities, Scope, Limitations, and Consequences, arXiv:2604.10332, 2026.

[11] OpenAI, GPT-4.1 Model Documentation, OpenAI API Docs. https://developers.openai.com/api/docs/models/gpt-4.1. Accessed: 2026-08-17.

[12] OpenAI, Introducing GPT-5, 2025. https://openai.com/index/introducing-gpt-5/. Accessed: 2026-08-17.

[13] Microsoft, GPT-5 vs GPT-4.1: Choosing the Right Model for Your Use Case, Microsoft Foundry Docs. https://learn.microsoft.com/en-us/azure/foundry/foundry-models/how-to/model-choice-guide. Accessed: 2026-08-17.

[14] Y. Wang, J. Huang, L. Du, Y. Guo, Y. Liu, and R. Wang, Evaluating Large Language Models as Raters in Large-Scale Writing Assessments: A Psychometric Framework for Reliability and Validity, Computers and Education: Artificial Intelligence, vol. 9, 100481, 2025.

[15] B. Braun and M. Forell, (Towards) Scalable Reliable Automated Evaluation with Large Language Models, arXiv:2607.28282, 2026.

[16] J. P. Kim, Augmenting Human Evaluation with LLM Judges: How Many Human Reviews Do You Need?, arXiv:2605.16354, 2026.

[17] P. Bhavsar, LLM-as-a-Judge vs Human Evaluation: When to Use Each (And Why Elite Teams Use Both), Galileo AI Blog, 2026. https://galileo.ai/blog/llm-as-a-judge-vs-human-evaluation. Accessed: 2026-08-18.

[18] A. Purwar and A. Choudhary, i-LAVA: Insights on Low Latency Voice-2-Voice Architecture for Agents, arxiv:2509.20971, 2025.

[19] A. Purwar and A. Choudhary, FOCAL: A Novel Benchmarking Technique for Multi-modal Agents, arxiv:2601.07367, 2026

APPENDIX

- **Critical Field Accuracy (CFA)**: Accuracy on error-sensitive entities (e.g., order ID, phone number, plan identifier) whose incorrect capture can invalidate task success.
- **Confirmation Precision (CP)**: The fraction of requested clarifications/confirmations that were actually necessary; a low CP indicates the agent over-clarifies when context was already sufficient.
- **Confirmation Recall (CR)**: The fraction of required clarifications/confirmations that were actually requested by the agent; a low CR means the agent proceeded on ambiguous input without asking.
- **Safety Recall (SR)**: Consistency with which the agent requests confirmation when required (e.g., under low ASR confidence or ambiguous intent), computed as confirmations requested over confirmation-required cases.
- **Irreversible Action Safety (IAS)**: The proportion of high-risk, irreversible actions (cancellations, charges, plan changes) that were executed only after explicit user confirmation; an IAS below 1.0 flags a critical safety failure.
- **Recovery Turn Count (RTC)**: The average number of conversational turns needed to recover from an error, covering ASR misrecognitions, tool failures, and incorrect agent actions.
- **Turn Efficiency (TE)**: The ratio of the optimal number of turns to the actual number of turns taken to complete a task, with values closer to 1.0 indicating efficient resolution.
- **User Experience Score (UES)**: A count of user repetitions, corrections, or restatements (e.g., re-spelling a name); high UES signals poor user experience even when the task ultimately succeeds.
- **ASR-Robust Goal Achievement (ARGA)**: The probability of achieving the task goal given that an ASR error occurred, isolating the agent's recovery capability from raw transcription accuracy.
- **Error Recovery Rate (ERR)**: The proportion of all detected errors, across ASR misrecognitions, tool failures, and wrong agent actions, that were successfully recovered via clarification, retry, or undo.