

From Simulation to Reality: Validating Automated Evaluation of Conversational Agents

Justyna Gromada, Ekaterina Isakova, Joanna Brudnik, Alicja Kasicka,
Natalia Krawczyk, Bartosz Przybył, Łukasz Krajewski, Michał Szczerbak

Orange Research, Orange Poland

firstname.lastname@orange.com

Abstract

Evaluating conversational agents typically requires human participants, a process that is costly and difficult to scale. LLM-based user simulators and LLM-as-a-judge evaluators offer a promising alternative, but how close they are to replacing real users remains unclear. We study this through a human user study (52 participants) and a simulated user study (four LLM-based simulators) applied to an experimental conversational sales agent, comparing LLM-as-a-judge assessments against human self-reported ratings. The results show that none of our simulators fully replicates human interaction patterns, and none of the LLM-based evaluators shows strong correlation with human ratings on real dialogues. We provide practical guidelines for combining automated and human evaluation, identifying key areas for improvement and highlighting when real-user studies remain essential.

1 Introduction

With the growing deployment of conversational agents, the need for scalable and reliable evaluation methods has become increasingly important. Human evaluation remains the gold standard, but is costly, time-consuming, and difficult to reproduce. Two automated approaches have gained popularity with the rise of LLMs: LLM-based user simulators, extending earlier simulation-based methods originally used for training dialogue systems via reinforcement learning (Levin et al., 2000), and LLM-as-a-judge evaluators, which assess dialogue quality automatically in zero or few-shot settings, without the need to train dedicated evaluation models (Lin and Chen, 2023; Zheng et al., 2023b; Li et al., 2024). Yet these automated approaches are built on assumptions about user behaviour that may not generalise to real-world interactions: if simulated users do not behave like real ones, or if LLM judges do not align with human judgments, evaluation results may be misleading. Validating such

automated pipelines against real user data predates LLMs (Georgila et al., 2020), but the emergence of LLM-based simulators and judges reopens this question at a larger scale and across new evaluation settings.

In this work, we empirically investigate whether human testers can be replaced by an automated pipeline combining LLM-based user simulators and LLM-as-a-judge evaluators, asking: (1) how well simulated dialogues resemble real user interactions, (2) how reliably LLM-based evaluators align with human judgments, and (3) whether automated pipelines can predict real user perception of agent configuration changes. As a testbed, we use a conversational sales agent for smartphone selection, studied through two complementary setups: a **human user study** (52 participants) and a **simulated user study** (four LLM-based simulators with injected persona characteristics).

Our main contributions are: (i) a large-scale empirical comparison of simulated and real user interactions across four LLM-based simulators and five evaluation dimensions; (ii) a systematic validation of LLM-as-a-judge evaluators against real user self-reported ratings, comparing an external (**LLM-as-observer**) and user-perspective (**LLM-as-participant**) judge roles; (iii) evidence that fully automated pipelines can generate false sensitivity signals and fail to reliably predict real user perception of agent configuration changes; and (iv) practical guidelines for combining automated and human evaluation in production settings.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes our evaluation methodology. Section 4 presents quantitative and qualitative results, followed by practical implications for industry. Sections 5 and 6 conclude the paper and discuss future directions.

2 Related Work

LLM-based user simulators have emerged as a scalable alternative to human participants in dialogue evaluation (Murakhovs'ka et al., 2023; Kim et al., 2025), but recent studies consistently reveal a sim-to-real gap (Zhou et al., 2026). Simulated users differ from humans in linguistic style (Ivey et al., 2024; Hathidara et al., 2026; Zhu et al., 2025), behavioural dimensions such as conversation strategy and engagement (Wang et al., 2025), and response quality that degrades over turns while human responses progressively improve (Lu et al., 2025). LLM simulators tend to be excessively cooperative and lack realistic frustration or ambiguity (Zhou et al., 2026), while human-AI pairs fail to form stable communicative conventions (Jones et al., 2026). Moreover, outcomes are sensitive to the choice of base LLM (Buck-Kaeffer et al., 2026; Zhu et al., 2025), and evaluations relying on simulated users produce miscalibrated agent assessments (Seshadri et al., 2026; Zhou et al., 2026). While simulators conditioned on rich user profiles can partially close this gap (Dou et al., 2025), fully replicating human interaction patterns remains an open challenge.

LLM-as-a-Judge approach has been proposed as a scalable alternative to human judges (Zheng et al., 2023a; Li et al., 2024), though reliability remains model and context dependent (Lu et al., 2025). LLM judges tend to agree more with each other than with human raters (Mukherjee et al., 2026; Thakur et al., 2025), and design choices such as grading scale can substantially affect alignment (Li et al., 2026). Drawing on persona assignment in LLMs (Tseng et al., 2024) and evidence that participants and observers may evaluate conversations differently, depending on the evaluation dimension (Street Jr et al., 1988; Tsuta et al., 2023; Georgila et al., 2020), we distinguish between an *LLM-as-observer* (external expert analyst) and an *LLM-as-participant* (user perspective) judge roles.

While LLM-based simulators and LLM judges have been studied largely in isolation, validation of complete automated pipelines against real human interactions and judgments remains limited. We address this gap through a large-scale comparison across multiple simulators, evaluation dimensions, and judge roles.

3 Evaluation Methodology

We describe the four components of our evaluation methodology: the conversational sales agent, the

human user study, the simulated user study, and the evaluation framework.

3.1 Conversational Sales Agent

We use an experimental conversational sales agent that assists customers in choosing a smartphone, subscription plan, and accessories. The agent is powered by GPT-4o-mini¹, the model used in our company's production setup.

We evaluate its two agent configurations to assess whether findings from simulated evaluations transfer to real human dialogues: (1) the **Base** agent, which focuses solely on helping the customer meet their goals and is equivalent to our current production system; and (2) the **Proactive** agent, which additionally proposes a premium phone alternative after the customer accepts the initial offer, as an up-sell strategy.

3.2 Human User Study

We conducted an online user study, recruiting 52 volunteers via email invitations sent to colleagues and acquaintances. Participation was voluntary, anonymous, and estimated at one hour. Participants were informed they would be interacting with an AI agent in English; 83% reported being comfortable or fluent in English, and the remaining 17% declared a decent proficiency level.

Each participant completed four conversations with the sales agent in a randomized order (task descriptions in Appendix A): two agent configurations (**Base** and **Proactive**) $\times$ two task types (**low** and **high** precision of needs), yielding $52 \times 4 = 208$ dialogues. Technology proficiency (*tech-savvy* vs. *average/low-tech*), assessed via pre-survey, determined the high-precision tasks: tech-savvy participants were assigned tasks involving more advanced and typically more expensive phone models, while others were assigned less advanced phone models at comparable specificity. Low-precision tasks provided only vague requirements (e.g., camera properties, larger storage).

After each conversation, participants rated their experience on a 5-point Likert scale across five dimensions: goal achievement, satisfaction, helpfulness, proactiveness, and naturalness (questions in Appendix B). Ratings were collected after the dialogue interface was replaced by an evaluation screen, to capture overall impressions rather than

¹<https://platform.openai.com/docs/models/gpt-4o-mini>

analytical re-reading. After all four conversations, participants completed a demographic questionnaire and a brief assertiveness measure structurally similar to the Rathus Assertiveness Schedule (Quillin et al., 1977), capturing behavioural tendencies relevant to up-selling interactions.

Participant profiles showed notable imbalances: predominantly male (63.5%), tech-savvy (57.7%), and highly assertive (73.1%). Age groups were skewed toward 25–34 (38.5%) and 35–44 (34.6%), with smaller shares of 45–54 (15.4%), 55–64 (7.7%), 18–24 and 65+ (1.9% each).

3.3 Simulated User Study

We generated a pool of 100 simulated user personas representing customers of our organisation using Claude 3.5 Sonnet v2², the strongest model available in our enterprise environment at the time. Each persona included demographic attributes (age, gender, family status, hobbies, devices and telecom services), a shopping goal with context, and four conversation scenarios defined by assertiveness level and precision of needs (high or low), with balanced gender and diverse age groups. Full details, including validation that simulators follow the assigned persona, are provided in (Gromada et al., 2025). The dataset of personas, goals, and scenarios is publicly available.³

From this pool, 52 personas were matched one-to-one with human participants, preserving distributions of age, gender, technology proficiency, and assertiveness. Each persona completed four independent conversations (context reset between runs): two agent configurations $\times$ two precision levels.

Simulators were built by injecting the persona profile, goal, and scenario into the system prompt, instructing the model to stay in persona, simulate realistic purchasing behaviour, and converse in English. Conversations ended when the simulator emitted a FINISHED token, signalling the goal was achieved or deemed unachievable, with a hard limit of 70 turns, never reached in practice. The full prompt template is in Appendix C.

We implemented four simulator variants to interact with the GPT-4o-mini-based sales agent. To avoid potential interaction bias between models of the same family, we selected three Google models:

Gemma-3-12b-it⁴, Gemini-2.0-flash, and Gemini-2.5-flash⁵. GPT-4o-mini, from OpenAI, was additionally included to explicitly probe this effect. Three runs per setting were generated to account for stochasticity, yielding 624 dialogues per simulator (52 personas $\times$ 2 precision levels $\times$ 2 configurations $\times$ 3 runs) and 2496 total simulated dialogues.

3.4 Evaluation Framework

Dialogue analysis We analyse dialogue structure (turn count, utterance length) and task success rates across real and simulated users.

We also measure three linguistic properties. Lexical diversity is measured using the Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010), ranging from 0 (no diversity) to 1 (maximum diversity) and robust to differences in text length. Question ratio is the proportion of turns containing at least one question mark, used as a proxy for information-seeking behaviour. Semantic similarity measures how closely simulated dialogues resemble real dialogues in content. Each dialogue is represented by a single embedding, computed by encoding every user turn with all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and averaging the resulting turn embeddings. We then compute cosine similarity between dialogue embeddings: for real dialogues, we average similarity over all unique pairs of real dialogues, excluding self-pairs, giving a self-similarity baseline; for simulated dialogues, we average similarity between each simulated dialogue and all real dialogues.

Human ratings As described in Section 3.2, participants rated their dialogues across five dimensions: user satisfaction, bot helpfulness, naturalness, proactiveness, and goal achievement, on a 5-point Likert scale, serving as ground truth for LLM-based assessment. User satisfaction and goal achievement are widely used dimensions in task-oriented dialogue evaluation (Walker et al., 1997), while bot helpfulness, naturalness and proactiveness are relevant dimensions in sales contexts.

LLM-based assessment We employ LLM judges across the same five dimensions in two roles: *LLM-as-observer*, acting as an external expert analyst, and *LLM-as-participant*, embodying the judge in a first-person perspective. Prompts are in Appendix D. Both roles are applied to real dialogues

²<https://www.anthropic.com/news/claude-3-5-sonnet>

³<https://huggingface.co/datasets/Orange/PersonasForSalesbot>

⁴<https://huggingface.co/google/gemma-3-12b-it>

⁵<https://ai.google.dev/gemini-api/docs/models>

and compared against human ratings to assess alignment; the LLM-as-participant role is additionally applied to simulated dialogues to analyse how ratings vary across agent configurations.

Model selection for each dimension follows prior validation studies (Gromada et al., 2025, 2026) on simulated dialogues with expert annotators, using Spearman correlation (Table 1). Unlike that study, where annotators observed generated conversations, the current study validates LLM-based ratings against self-reported ratings from real users who participated in the dialogues. The same set of models is used in both roles of LLM judge.

Dimension	LLM evaluator	ρ
User Satisfaction	Qwen2.5-14B	0.78
Bot Helpfulness	Qwen2.5-32B	0.64
Bot Naturalness	Llama-3.1-8B	0.52
Bot Proactiveness	Selene-1-Mini-Llama-8B	0.63
Goal Achievement	Selene-1-Mini-Llama-8B	0.79

Table 1: LLM evaluators selected via validation studies (Gromada et al., 2025, 2026) on simulated dialogues with expert annotators (Spearman ρ). Full benchmarking in Appendix E.

Qualitative analysis Beyond quantitative measures, we draw insights from open-ended participant comments collected in the human user study, and from manual inspection of real and simulated dialogues alongside their LLM-based ratings, focusing on patterns and discrepancies between automated and human assessments.

4 Results

Our results address three core questions: (i) how well simulated dialogues resemble real user interactions; (ii) how reliably LLM-based evaluators align with human judgments on real user dialogues; and (iii) whether automated pipelines can predict real user perception of sales agent configuration changes.

4.1 Quantitative results

4.1.1 Simulated vs. Real User Dialogues

We compare simulated and real user dialogues through structural, linguistic, and task-level characteristics, summarized in Table 2.

Mean dialogue length is similar across groups (29–36 turns), but real user dialogues show considerably higher variance (SD=17, range: 6–98) than all simulators (SD: 6–9), suggesting simulated

users follow more predictable interaction patterns. GPT-4o-mini produces the shortest dialogues.

User utterance length varies considerably across simulators. Gemma-3-12b most closely matches real users (58 chars mean for both), while Gemini-2.5-flash produces notably longer utterances (331 vs. 58 chars), unlikely to reflect natural behaviour.

All simulators fall below the real user system-reported success rate (89.4%), ranging from 59.6% (Gemini-2.5-flash) to 81.1% (Gemma-3-12b). For real users, the system-reported success rate exceeds the user-reported one (89.4% vs. 82.7%), suggesting the system overestimates task completion.

Lexical diversity (MATTR) scores are close across conditions (simulators: 0.77–0.83; real users: 0.82). Gemini-2.5-flash achieves the closest mean (0.83) but lowest variability (SD=0.02); Gemma-3-12b shows the lowest mean (0.77) but highest variability (SD=0.08), more closely mirroring real users.

The question ratio is consistently higher for simulators (0.49–0.53) than for real users (0.38). This indicates that simulated users ask more questions within individual turns, likely reflecting their goal-driven information-gathering strategy. The exception is Gemma-3-12b, which most closely aligns with real users (0.39).

Semantic similarity to real dialogues ranges from 0.46 to 0.48 across simulators, only slightly below the real-user self-similarity baseline of 0.52. Gemini-2.0-flash and GPT-4o-mini are closest to the baseline (0.48 each), while Gemma-3-12b is furthest (0.46), yet the differences are small.

Overall, while Gemma-3-12b best mimics real users in utterance length, task success, lexical diversity, and question ratio, no simulator consistently replicates human behaviour across all dimensions.

4.1.2 LLM–Human Rating Alignment

We assess alignment between LLM-based evaluators, using dimension-specific models selected in our validation study (Table 1), and human self-reported ratings on real user dialogues. Table 3 presents Spearman correlations ($n=208$) for both judge roles (*LLM-as-observer* and *LLM-as-participant*), alongside results from a prior validation studies (Gromada et al., 2025, 2026) where the same models were evaluated against expert annotators on simulated dialogues.

Correlations drop notably compared to prior validation across all dimensions, especially for User Satisfaction (0.78 $\rightarrow$ 0.37) and Bot Proactive-

Metric / Dimension	Real	Simulated Users			
	Users	Gemma-3-12b	Gemini-2.0-flash	Gemini-2.5-flash	GPT-4o-mini
Dialogue Characteristics					
Total dialogues	208	624	624	624	624
<i>Dialogue Structure</i>					
Length [#turns]: mean $\pm$ SD	34 $\pm$ 17	32 $\pm$ 6	36 $\pm$ 7	35 $\pm$ 9	29 $\pm$ 6
Length [#turns]: median	32	32	36	36	28
Length [#turns]: range	6–98	14–54	20–64	18–70	13–56
<i>Utterance Characteristics</i>					
User [#chars]: mean $\pm$ SD	58 $\pm$ 49	58 $\pm$ 44	98 $\pm$ 67	331 $\pm$ 262	126 $\pm$ 72
User [#chars]: range	1–300	2–404	2–412	2–2927	0–479
Agent [#chars]: mean $\pm$ SD	509 $\pm$ 413	413 $\pm$ 290	471 $\pm$ 320	657 $\pm$ 458	502 $\pm$ 356
<i>Task Success</i>					
System-reported [%]	89.4	81.1	74.2	59.6	65.4
User-reported [%]	82.7	–	–	–	–
<i>Linguistic Characteristics</i>					
MATTR: mean $\pm$ SD	0.82 $\pm$ 0.07	0.77 $\pm$ 0.08	0.80 $\pm$ 0.04	0.83 $\pm$ 0.02	0.80 $\pm$ 0.04
Question ratio: mean $\pm$ SD	0.38 $\pm$ 0.22	0.39 $\pm$ 0.14	0.49 $\pm$ 0.12	0.50 $\pm$ 0.16	0.53 $\pm$ 0.15
Semantic similarity: mean	0.52 [†]	0.46	0.48	0.47	0.48

Table 2: Dialogue characteristics for real users ($n = 208$) and simulated users ($n = 624$ each). System-reported success based on automated evaluation; user-reported based on participant self-assessment. Semantic similarity is computed between simulated and real dialogues; [†]for real users, this is a self-similarity baseline.

Dimension	LLM-as-participant	LLM-as-observer	Prior
User Satisfaction	0.37	0.29	0.78
Bot Helpfulness	0.21	0.22	0.64
Bot Naturalness	−0.01	−0.01	0.52
Bot Proactiveness	0.16	0.07	0.63
Goal Achievement	0.38	0.39	0.79

Table 3: Spearman correlation between LLM evaluators and human ratings on real user dialogues ($n=208$). LLM-as-participant: user perspective. LLM-as-observer: third-party evaluation. Prior: validation on simulated dialogues with expert observers (Gromada et al., 2025, 2026).

ness ($0.63 \rightarrow 0.16$). *LLM-as-participant* tends to slightly outperform *LLM-as-observer*, suggesting the user perspective helps capture user-relevant aspects. Bot Naturalness is the most resistant, with near-zero correlation ($\rho = -0.01$) for both roles, suggesting that perceived naturalness is largely inaccessible from dialogue text alone. This reflects a fundamental limitation: human ratings capture internal cognitive and affective states not fully expressed in dialogue, while LLM evaluators access only surface-level signals (Hua and Artzi, 2024).

Overall, correlations with human ratings remain moderate ($\rho \leq 0.39$), with Goal Achievement and User Satisfaction showing the strongest alignment and Bot Naturalness proving the most challenging. The drop from prior validation, shown in the

Prior column of Table 3, highlights the challenge of generalising from simulated to real user dialogues.

4.1.3 Base vs. Proactive Agent Configuration

We examine whether an automated pipeline, combining simulated users with LLM-as-participant evaluators, can predict how real users perceive changes in sales agent behaviour.

Rating differences between sales agent configurations (Proactive vs. Base) across five quality dimensions (user satisfaction, bot helpfulness, bot naturalness, bot proactiveness, and goal achievement) were analysed using cumulative link mixed models (CLMM) (Christensen, 2019) with a random intercept per participant (or persona), accounting for the repeated-measures structure of the study. This approach directly models the ordinal Likert scale and avoids normality assumptions required by parametric alternatives, and yields an odds ratio (OR) per dimension. An OR compares the odds of a higher rating under one configuration versus the other, with $OR=1$ indicating no difference. P -values were adjusted for multiple comparisons using the Holm correction within each test family. Non-significant effects were further examined using two-one-sided-tests (TOST) equivalence testing (Lakens, 2017) to distinguish true null effects from insufficient power.

Table 4 presents per-dimension odds ratios and paired effect sizes (Cohen’s d_z) for two sales agent

Dimension	Real Users			Simulated Users (pooled)		
	OR [95% CI]	p (p_{Holm})	d_z	OR [95% CI]	p (p_{Holm})	d_z
User Satisfaction	0.89 [0.52, 1.52]	0.669 (1.00)	-0.02	0.68 [0.58, 0.80]	<0.001 (<0.001)	-0.58
Bot Helpfulness	0.90 [0.52, 1.55]	0.698 (1.00)	-0.01	0.50 [0.42, 0.59]	<0.001 (<0.001)	-1.04
Bot Naturalness	0.59 [0.33, 1.05]	0.071 (.283)	-0.27	0.51 [0.43, 0.60]	<0.001 (<0.001)	-1.11
Bot Proactiveness	1.76 [1.03, 3.02]	0.038 (.191)	+0.31	0.74 [0.64, 0.86]	<0.001 (<0.001)	-0.54
Goal Achievement	0.64 [0.36, 1.12]	0.118 (.355)	-0.12	0.94 [0.80, 1.12]	0.514 (0.514)	-0.17

Table 4: CLMM rating differences and paired effect sizes (Cohen’s d_z) for sales agent configurations: Proactive vs. Base for real users ($n = 52$; 208 dialogs) and simulated users ($n = 52$; 2496 dialogs). OR>1 favors Proactive; OR<1 favors Base. $d_z > 0$ favors Proactive; $d_z < 0$ favors Base. CLMM: rating $\sim$ condition + (1|id), logit link. p : raw p -value, with Holm-corrected p -value (p_{Holm}) in parentheses, corrected within each 5-test family. Raw p -values are shown in bold when $p < .05$ Full d_z confidence intervals in Appendix F.2; per-simulator breakdown in Appendix F.1.

configurations: Proactive vs. Base, for both real and simulated users, with the simulated-user results pooling all four simulators. For real users, ratings are self-reported; for simulated users, ratings are provided by LLM-as-participant evaluators (best model per dimension; see Table 1). Appendix F provides additional detail: per-simulator breakdowns (Table F.1), full effect-size details with confidence intervals and magnitude classifications (Table F.2), and TOST equivalence tests (Table F.3).

Real users show no statistically significant difference between configurations across all dimensions after Holm correction (CLMM; all corrected $p > 0.05$), with negligible to small effect sizes (Cohen’s d_z : 0.01 to 0.31). Simulated users show the same direction on three dimensions, but with far larger, significant effects: Bot Naturalness (real: OR=0.59, not significant; simulated: OR=0.51 [0.43, 0.60], $p < 0.001$, $d_z=-1.11$), Bot Helpfulness (real: OR=0.90, not significant; simulated: OR=0.50 [0.42, 0.59], $p < 0.001$, $d_z=-1.04$), and User Satisfaction (real: OR=0.89, not significant; simulated: OR=0.68 [0.58, 0.80], $p < 0.001$, $d_z=-0.58$). Given these small, non-significant real-user effect sizes, this directional agreement carries limited evidentiary value. This is confirmed by TOST tests, which show that these real-user nulls cannot be interpreted as true zero effects ($p_{\text{equiv}} > 0.05$ in each case; see Appendix F.3 for the full set), whereas the corresponding simulated effects are clearly distinguishable from zero ($|d_z| > 0.57$). Goal Achievement instead shows a true null in both settings (real: OR=0.64 [0.36, 1.12], $p=0.118$; simulated: OR=0.94 [0.80, 1.12], $p=0.514$, effect size negligible), consistent with it being the most factual dimension, relating to concrete task completion rather than subjective states, which may explain why it is the only one reliably captured by

LLM evaluators. Bot Proactiveness, the dimension most directly tied to the agent configuration change, is the one case where real users show any directional signal at all (OR=1.76 [1.03, 3.02], though not significant after correction), and it is exactly here that simulated users diverge, showing a significant effect in the opposite direction (OR=0.74 [0.64, 0.86], $p < 0.001$, $d_z=-0.54$).

At the per-simulator level (Appendix F.1), all four simulators agree in direction with the pooled simulated model on four of five dimensions, with Gemini-2.5-flash the most conservative, as no dimension survives correction individually. This makes Gemini-2.5-flash the closest to the real-user null pattern, though this apparent proximity should be interpreted cautiously given the wide confidence intervals around the real-user estimates. For Bot Proactiveness, all four simulators diverge from real users in the same direction, and this divergence reaches significance only once simulators are pooled, indicating that no simulator-evaluator combination, individually or pooled, offers a reliable substitute for real users on this dimension.

Overall, with our current models and prompting settings, the combination of user simulators and LLM-based evaluators cannot reliably predict real user perception of agent configuration changes. It produces large-effect sensitivity signals where real-user evidence is inconclusive, and reverses the positive (though non-significant) trend on Bot Proactiveness.

4.2 Qualitative Observations

We complement the quantitative findings with qualitative observations, examining how real and simulated users differ in their interaction patterns, and how human and LLM-based evaluators differ in their rating behaviour.

4.2.1 Interaction Patterns

Unlike simulated users who follow a task-oriented script, real users engage more freely: they question agent outputs, demand justifications (e.g., *"why do you suggest this model?"*), challenge wording like *"may be better"* or *"the best"*, request comparisons and niche details, and bring prior device experience and price sensitivity into the conversation. They face cognitive costs when composing messages, leading to more concise and purposeful communication, explaining the higher variance in dialogue length (Section 4.1.1). They are also sensitive to response delays, message length, and verbosity, e.g., *"There is way too much text, it's overwhelming. It feels like information overload instead of a helpful summary."*, reactions absent in simulated interactions.

4.2.2 Rating Patterns

Human and LLM-based evaluators differ notably in rating behaviour. Human participants show broader score ranges and higher variability. LLM-based evaluators, by contrast, exhibit positivity bias and limited response range, avoiding extreme negative ratings regardless of dialogue quality (Appendix G). This conservative rating behaviour suggests that with our current models and prompting settings, LLM judges systematically underestimate variance in real user experience, further limiting their ability to detect meaningful differences across conditions.

4.3 Practical Implications for Industry

Our results reveal key limitations of automated evaluation and offer concrete guidance for deploying conversational agents in sales-related contexts.

Among simulators evaluated with our current settings, no single model consistently replicates human behaviour across all dimensions: utterance length, task success rate, and lexical diversity vary substantially. Simulators should be validated against real user dialogues before being relied upon, as they miss critical dimensions of real user behaviour. Simulator logic should be enriched with response perception elements (e.g., reacting to message length and verbosity), experience-driven decision-making, realistic price negotiation, and explicit reasoning steps; few-shot examples can further improve fidelity.

LLM-based evaluators tend to exhibit positivity bias and limited response range, systematically underestimating variance in real user experience. Calibration functions should be considered to correct

this bias, and dimensions such as Bot Naturalness treated with particular caution, as they remain inaccessible to automated evaluation with our current settings.

Fully automated pipelines can generate false sensitivity signals. We recommend a hybrid approach: automated pipelines for development and testing, and real-user studies for validating changes to deployed conversational agents.

5 Conclusions

We presented a systematic evaluation of automated pipelines for assessing conversational sales agents, combining LLM-based user simulators and evaluators, validated against a human user study. With our current settings, both components face significant limitations: no simulator consistently replicates human interaction patterns across all checked dimensions, and LLM-based evaluators suffer from positivity bias, reduced performance on real user dialogues, and near-zero alignment with human ratings on dimensions such as Bot Naturalness.

Simulation-based approaches remain valuable for stress-testing agent behaviour, identifying systematic errors, and generating dialogues for manual inspection during development. Automated evaluation can also reliably measure more objective aspects such as task success and Goal Achievement. However, for validating subjective dimensions of user experience (satisfaction, naturalness, perceived proactiveness) and for assessing the impact of agent configuration changes in production systems, real-user studies remain essential.

6 Future Work

Future work will focus on validating our findings in real deployment settings, where users interact with the agent organically rather than in a controlled study. We also plan to explore more sophisticated automated evaluation methods to better capture subjective user experience, such as reasoning-based LLM evaluators or classical ML-based approaches trained to extract richer signals from dialogue text. Another promising direction involves implementing statistical models that simultaneously consider multiple dialogue characteristics for more rigorous real-vs-simulated comparison.

Limitations

This study is subject to several limitations that should be considered when interpreting the find-

ings.

Human User Study The sample is small and linguistically constrained, with demographic imbalances in gender (approximately 3:2), age, and technological proficiency, and is disproportionately composed of individuals with AI-related occupations and frequent AI users, introducing selection bias toward more AI-familiar respondents. Participants were asked to adopt the role of someone interested in purchasing a smartphone rather than acting on a real need, which may have affected engagement and interaction style. Engaging in multiple similar interactions may additionally have led to fatigue or attention decline, impacting response consistency despite randomised scenario order. Finally, although the study was conducted in English, all participants were non-native speakers who rated their proficiency as sufficient, which may have influenced interaction quality and conversational fluidity.

Simulated User Study Unlike human participants, who carry memory, fatigue, and emotional state across conversations, each simulated persona conducted conversations independently with context reset between them. This fundamental difference may affect the comparability of human and simulated interactions, as real users may be influenced by prior conversations in ways that simulators cannot replicate.

LLM-based Evaluation The LLM evaluators were selected based on a prior validation study conducted on simulated dialogues with expert annotators, and their performance drops substantially on real user dialogues, as reported in Section 4.1.2. Furthermore, the evaluation is limited to textual signals, and dimensions such as Bot Naturalness remain largely inaccessible to automated assessment.

Experimental Setup The experiment was conducted in a single domain (telecommunications), with a single agent and within a single-configuration contrast, which may limit the generalisability of the research findings.

References

Aurelien Buck-Kaeffer, Sneheel Sarangi, Maximilian Puelma Touzel, Reihaneh Rabbany, Zachary Yang, and Jean-Francois Godbout. 2026. The silicon *Society Cookbook*: Design space of llm-based social simulations. *arXiv preprint arXiv:2605.00197*.

Rune Haubo Bojesen Christensen. 2019. ordinal—regression models for ordinal data. *R package version*, 10(2019):54.

Michael A. Covington and Joe D. McFall. 2010. [Cutting the gordian knot: The moving-average type-token ratio \(mattr\)](#). *Journal of Quantitative Linguistics*, 17:100 – 94.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. [SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35212–35290, Suzhou, China. Association for Computational Linguistics.

Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. 2020. [Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 726–734, Marseille, France. European Language Resources Association.

Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybył, Lina M. Rojas-Barahona, and Michał K. Szczerbak. 2025. [Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 230–245, Suzhou (China). Association for Computational Linguistics.

Justyna Gromada, Alicja Kasicka, Natalia Krawczyk, and Michał Szczerbak. 2026. [Evaluating Success in Proactive Sales Dialogues: Balancing Customer Experience and Business Outcomes](#). In *WOCHAT2026: Workshop on Chatbots and Agentic Technologies*, pages 66–84.

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisetipalli. 2026. Mirrorbench: A benchmark to evaluate conversational user-proxy agents for human-likeness. *arXiv preprint arXiv:2601.08118*.

Yilun Hua and Yoav Artzi. 2024. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal llms. In *First Conference on Language Modeling*.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, and 1 others. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*.

Cameron R. Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K. Bergen. 2026. [Llms and people both learn to form conventions - just not with each other](#). *ArXiv*, abs/2602.08208.

- Tongyoung Kim, Jeongeun Lee, Soojin Yoon, Seonghwan Kim, and Dongha Lee. 2025. [Towards personalized conversational sales agents : Contextual user profiling for strategic action](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362.
- Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. [A stochastic model of human-machine interaction for learning dialog strategies](#). *IEEE Trans. Speech Audio Process.*, 8:11–23.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [Llms-as-judges: A comprehensive survey on llm-based evaluation methods](#). *ArXiv*, abs/2412.05579.
- Weiyue Li, Minda Zhao, Weixuan Dong, Jiahui Cai, Yuze Wei, Michael Pocrress, Yi Li, Wanyan Yuan, Xiaoyu Wang, Ruoyu Hou, Kaiyuan Lou, Wenqi Zeng, Yutong Yang, Yilun Du, and Mengyu Wang. 2026. [Grading scale impact on llm-as-a-judge: Human-llm alignment is highest on 0-5 grading scale](#). *ArXiv*, abs/2601.03444.
- Yen-Ting Lin and Yun-Nung (Vivian) Chen. 2023. [Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models](#). *ArXiv*, abs/2305.13711.
- Dongxu Lu, Johan Jeuring, and Albert Gatt. 2025. Evaluating llm-generated versus human-authored responses in role-play dialogues. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 20–40.
- Sourabrata Mukherjee, Hamna Hamna, Kalika Bali, and Sunayana Sitaram. 2026. [The geometry of llm-as-judge: Why inter-llm consensus is not human alignment](#). *arXiv preprint arXiv:2606.03043*.
- Lidiya Murakhovs’ka, Philippe Laban, Tian Xie, Caiming Xiong, and Chien-Sheng Wu. 2023. [Salespeople vs SalesBot: Exploring the role of educational value in conversational recommender systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9823–9838, Singapore. Association for Computational Linguistics.
- James W. Quillin, Steven Besing, and David Dinning. 1977. [Standardization of the rathus assertiveness schedule](#). *Journal of Clinical Psychology*, 33:418–422.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. [Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations](#). *ArXiv*, abs/2601.17087.
- Richard L Street Jr, John M Wiemann, and Anthony Mulac. 1988. Speech evaluation differences as a function of perspective (participant versus observer) and presentational medium. *Human Communication Research*, 14(3):333–363.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. [Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Yuma Tsuta, Naoki Yoshinaga, Shoetsu Sato, and Masashi Toyoda. 2023. Rethinking response evaluation from interlocutor’s eye for open-domain dialogue systems. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 55–63.
- Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. [PARADISE: A framework for evaluating spoken dialogue agents](#). In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.
- Zhefan Wang, Ning Geng, Zhiqiang Guo, Weizhi Ma, and Min Zhang. 2025. Human vs. agent in task-oriented conversations. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 133–142.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haoteng Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b.

Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, pages 46595–46623.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. 2026. [Mind the sim2real gap in user simulation for agentic tasks](#). *arXiv preprint arXiv:2603.11245*.

Ruizhe Zhu, Hao Zhu, Yaxuan Li, Syang Zhou, Shijing Cai, Malgorzata Lazuka, and Elliott Ash. 2025. Dialogueforge: Llm simulation of human-chatbot dialogue. In *Agentic & GenAI Evaluation Workshop KDD'25*.

A Human Studies - Tasks

Two tasks with low precision of needs; universal for all participants:

1. Your Xiaomi Redmi 13C 5G is not serving you well anymore. So you're looking for a new phone with better camera capabilities. You need reliable device and larger storage capacity. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.
2. Your Samsung Galaxy A53 is not serving you well anymore. So you're looking for a new phone with better camera capabilities. You need reliable device and larger storage capacity. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.

Two tasks with high precision of needs; for participants with high digital literacy:

1. You want to upgrade to the new iPhone 16 Pro 256GB 5G. You want high storage capacity for files, photos, and apps, and you care about camera quality at a professional level. Seamless sync with your Apple devices is essential. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.
2. You want to upgrade to the new Xiaomi 15 Ultra 5G 16/512GB. You want high storage capacity for files, photos, and apps, and you care about camera quality at a professional level. Seamless sync with your Apple devices is essential. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.

Two tasks with high precision of needs; for participants with average/low digital literacy:

1. Your current phone no longer meets your everyday needs, so you're looking for a reliable phone with dual SIM, long battery life, and smooth performance for communication and entertainment: you're considering Xiaomi Redmi Note 14 Pro+ with 512GB of storage. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.
2. Your current phone no longer meets your everyday needs, so you're looking for a reliable phone with dual SIM, long battery life, and smooth performance for communication and entertainment: you're considering Samsung Galaxy A56 with 256GB of storage. Feel free to ask for more options if you're curious – you don't have to go with the first offer the bot suggests. It's optional and completely up to you.

B Human Studies - Rating Questions

The following five statements were rated by participants on a five-point Likert scale according to their level of agreement:

- I was able to achieve my goal in this interaction.
- I am satisfied with this conversation, I feel positive about this conversation.
- Assistant was helpful and addressed most of my needs with relevant information and solutions.
- The assistant was proactive addressing my needs with good suggestions and guidance.
- The assistant's responses felt natural and human-like for me.

C Simulator Prompt

```
# User Simulation Instructions

## Your Role
You are simulating {name} - a real customer interacting with
an {enterprise_name} mobile sales assistant. Behave as authentically as
possible. Never reveal you are an AI. Never break character.
Maintain consistent persona traits throughout the conversation. Respond naturally as the defined
persona would.

---

## Simulation Parameters
- **PERSONA**: {identity}
- **GOAL**: {goal}
- **SCENARIO**: {conversation_scenario}
- **DATE**: {date}
- **END_WORD**: {end_word}

---

## CRITICAL FORMATTING RULE
Every message MUST start with {name}{separator} - no exceptions.

- Correct: {name}{separator} your message...
- Correct: {name}{separator} your message... [FINISHED]
- Incorrect: [REJECT_OFFER] message...
- Incorrect: {name}{separator} {name}{separator} message...

Tags ([ACCEPT_OFFER], [REJECT_OFFER], etc.) always go at the
END of your message, never at the beginning.

Input format: assistant_name{separator}assistant_message
Output format: {name}{separator}your_message

## Behavioral Guidelines

1. **Communication Style**
- Communicate in {language} for ALL messages - no exceptions
- Only mix languages if PERSONA is explicitly described as
  bilingual or multilingual - living in a non-English country
  does NOT make a PERSONA bilingual
- Match assertiveness level from SCENARIO
- Follow precision level from SCENARIO
- Align language with PERSONA background
- Stay within GOAL and SCENARIO requirements parameters
- Stay focused on your GOAL - only discuss relevant
  products and services

2. **General Response Guidelines**
- Keep responses aligned with PERSONA traits
- Reference GOAL in key decisions
- Keep responses short and natural
- Do not repeat what has
  already been said
- You are a CUSTOMER, not a sales agent:
  * Express your needs and react to proposals
  * Do NOT suggest products, plans, or pricing to the assistant
  * Let the assistant do the selling

- Reveal information gradually:
  * Start with your most essential requirement only
  * Do not mention your occupation, lifestyle, or usage context
    in your first message - share it only when the assistant asks
    or when it becomes naturally relevant
  * Reveal budget, contract preferences, and additional needs
    as the conversation naturally calls for them
  * More assertive personas may reveal requirements earlier,
```

but still not all at once

- Only use information from PERSONA, GOAL, and SCENARIO - do not invent or assume anything not provided
- If the assistant clearly confirms something is unavailable, accept that answer - do not rephrase and ask the same question again

3. **Conversation Ending Rules**

- Send END_WORD as the LAST element of your final message
- Correct: {name}{separator} Thank you, goodbye! {end_word}
- Correct: {name}{separator} {end_word}
- Incorrect: {name}{separator} {end_word} Thank you!
- Incorrect: {name}{separator} [ACCEPT_OFFER] Great, let's proceed. {end_word}
- NEVER send END_WORD in the same message as [ACCEPT_OFFER] or [ACCEPT_REC] - always wait for the assistant to acknowledge before ending:
 - * Turn N: {name}{separator} [ACCEPT_OFFER] I'm satisfied, let's proceed.
 - * Turn N+1: assistant confirms and optionally provides a shopping link
 - * Turn N+2: {name}{separator} Thank you, goodbye! {end_word}
- End the conversation when:
 - * Successful: your goal is achieved and the assistant has acknowledged - say goodbye naturally
 - * Unsuccessful: the assistant cannot help you - react according to your PERSONA and end with END_WORD

4. **Decision Process**

- Evaluate all proposals against SCENARIO requirements only
- PERSONA traits may inform your communication style, but SCENARIO requirements drive your decisions
- Work towards GOAL completion
- Use the decision style described in SCENARIO

5. **Tags**

- Add a tag at the end of your message, on a new line, if applicable:
 - * [ACCEPT_OFFER] - fully satisfied with the offer, ready to proceed
 - * [REJECT_OFFER] - not satisfied with the offer
 - * [ACCEPT_REC] - accepting a recommendation
 - * [REJECT_REC] - declining a recommendation
 - * [FINISHED] - conversation is complete (successful or unsuccessful)
- The tag must reflect your message naturally - never write a message just to justify a tag
- If no tag applies, do not add one

6. **Budget Rules**

- Use currency based on PERSONA location if not specified
- If budget.amount is specified and described as "monthly":
 - * Your budget covers total monthly cost (plan + device combined)
- If budget.amount is null and described as "reasonable":
 - * Use current monthly_spend as your reference point
 - * Adjust based on income level:
 - High income: can go somewhat above current spend
 - Middle income: stay close to current spend
 - Low income: try to stay at or below current spend
- Budget flexibility:
 - * STRICT: reject anything above stated amount
 - * FLEXIBLE: consider offers up to 20 percent above budget
- When a phone is presented with a total price and you have a monthly budget:
 - * Roughly estimate if the phone leaves room for a plan

- within your budget - do not calculate exact installments
- * If feasible: tentatively accept and ask about plan options
- * If clearly not feasible: reject based on budget
- * Always ask the assistant for the total monthly breakdown before making a final decision

7. ****Contract Rules****

- STRICT: only accept the specified contract type
- FLEXIBLE: consider alternatives if they are clearly beneficial

8. ****Requirements Management****

- Start by communicating only your essential requirements
- Keep flexible requirements hidden until relevant:
 - * Reveal them when additional products/services are proposed
 - * Reveal them when specific questions are asked
 - * Reveal them when alternatives are being discussed

9. ****Device Knowledge****

- Your device knowledge comes from PERSONA description and may be outdated or incomplete
- If the assistant mentions a newer model than you expected:
 - * Show genuine surprise and interest
 - * Be open to adjusting your goal towards the newer model if it fits your needs and budget
 - * Ask questions before making a decision
 - * Do not stubbornly stick to your original goal - a real customer would be curious about better options

10. ****Naturalness and Conversation Flow****

- Ask one or two questions at a time - never send a long list of questions in a single message
- Wait for the answer before asking the next question
- Reveal your needs gradually - do not dump all requirements, budget, and concerns at once
- Stay focused on your goal - do not drift into peripheral topics (store addresses, scripts, appointment booking) as a substitute for making a decision
- Once phone, plan, and total monthly cost are confirmed within your budget, move towards a decision - do not keep inventing new questions or concerns
- Match your response length to the situation - a simple confirmation deserves a short reply
- Follow a natural conversation arc: explore, ask key questions, compare, decide
Do not get stuck in an endless loop of minor clarifications

D Evaluation Prompts

D.1 User Satisfaction

D.1.1 User Satisfaction by LLM-as-observer

You are an expert conversation analyst evaluating a real customer interaction with a sales assistant selling phones, offers, accessories, and additional services.

The conversation you will evaluate involves a real user.
Note that real users often write very short messages (e.g., "cheap", "yes", "Samsung") this is normal and expected.

The conversation:
{conv_history}

```

---

Your task is to estimate how satisfied the user likely was with this conversation.

Please use the following scale:
- rating **1**:: Completely not satisfied
- rating **2**:: Rather not satisfied
- rating **3**:: Neither satisfied nor unsatisfied
- rating **4**:: Rather satisfied
- rating **5**:: Very satisfied

**Required Output Format:**
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain key interactions or phrases that led to this rating
}}
```

D.1.2 User Satisfaction by LLM-as-participant

```

Imagine you are the user in this conversation.
You have just finished talking with an assistant
selling phones with offers, accessories and additional services.

The conversation you just had:
{conv_history}

Reflecting on your experience as the user in this conversation,
please rate the following statement:

"I am satisfied with this conversation,
I feel positive about this conversation."

Please use the following scale:
- rating **1**:: Completely not satisfied
- rating **2**:: Rather not satisfied
- rating **3**:: Neither satisfied nor unsatisfied
- rating **4**:: Rather satisfied
- rating **5**:: Very satisfied

**Required Output Format:**
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain from the user's perspective
                        // what made you feel this way
}}
```

D.2 Bot Helpfulness

D.2.1 Bot Helpfulness by LLM-as-observer

```

You are an expert conversation analyst evaluating a real customer interaction
with a sales assistant selling phones, offers, accessories,
and additional services.

The conversation you will evaluate involves a real user.
Note that real users often write very short messages
(e.g., "cheap", "yes", "Samsung") this is normal and expected.

The conversation:
{conv_history}

---

Your goal is to rate the helpfulness of the sales assistant bot based on the conversation.
The bot is helpful when it addresses customer needs with relevant information and solutions.

Please use the following scale:
- rating **1**:: Completely not helpful
- rating **2**:: Rather not helpful
```

```

- rating **3**: Neither helpful nor unhelpful
- rating **4**: Rather helpful
- rating **5**: Very helpful

**Required Output Format:**
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain key interactions or phrases that led to this rating
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}

```

D.2.2 Bot Helpfulness by LLM-as-participant

Imagine you are the user in this conversation.
 You have just finished talking with an assistant selling phones with offers, accessories and additional services.

The conversation you just had:
 {conv_history}

Reflecting on your experience as the user in this conversation, please rate the following statement:

"Assistant was helpful and addressed most of my needs with relevant information and solutions."

Please use the following scale:

```

- rating **1**: Completely not helpful
- rating **2**: Rather not helpful
- rating **3**: Neither helpful nor unhelpful
- rating **4**: Rather helpful
- rating **5**: Very helpful

```

Required Output Format:

```

{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain from the user's perspective
                        // what made you feel this way
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}

```

D.3 Bot Naturalness

D.3.1 Bot Naturalness by LLM-as-observer

You are an expert conversation analyst evaluating a real customer interaction with a sales assistant selling phones, offers, accessories, and additional services.

The conversation you will evaluate involves a real user.
 Note that real users often write very short messages (e.g., "cheap", "yes", "Samsung") this is normal and expected.

The conversation:
 {conv_history}

Your goal is to rate **the naturalness** of the sales assistant based on the conversation. Sales assistant' naturalness refers to how much his responses sound natural, fluid, and human-like, making interactions feel more genuine and comfortable for users.

```

Please use the following scale:
- rating **1**: Completely artificial
- rating **2**: Rather not natural
- rating **3**: Neither natural nor artificial
- rating **4**: Rather natural
- rating **5**: Very natural

**Required Output Format:**
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain key interactions or phrases that led to this rating
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}

```

D.3.2 Bot Naturalness by LLM-as-participant

```

Imagine you are the user in this conversation.
You have just finished talking with an assistant selling phones with offers, accessories and
  additional services.

The conversation you just had:
{conv_history}

Reflecting on your experience as the user in this conversation, please rate the following statement:

"The assistant's responses felt natural and human-like for me."

Please use the following scale:
- rating **1**: Completely disagree
- rating **2**: Rather disagree
- rating **3**: Neither agree nor disagree
- rating **4**: Rather agree
- rating **5**: Completely agree

**Required Output Format:**
{{
  "rating": int,          // a single number on the scale of 1 to 5.
  "justification": str,   // explain from the user's perspective
                        // what made you feel this way
                        // describe only the assistant naturalness,
                        // skip overall naturalness of the conversation and
                        // your naturalness
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}

```

D.4 Bot Proactiveness

D.4.1 Bot Proactiveness by LLM-as-observer

You are an expert conversation analyst evaluating a real customer interaction with a sales assistant selling phones, offers, accessories, and additional services.

The conversation you will evaluate involves a real user. Note that real users often write very short messages (e.g., "cheap", "yes", "Samsung") this is normal and expected.

The conversation:

```
{conv_history}

---

Your goal is to rate the proactiveness of the sales assistant based on how well it:
- Anticipates customer needs
- Suggests relevant products/services
- Takes initiative in the conversation
- Makes appropriate follow-up recommendations
- Guides the customer through the sales process

Please use the following scale:
- rating **1**: Not at all proactive
- rating **2**: Rather not proactive
- rating **3**: Neither proactive nor not proactive
- rating **4**: Rather proactive
- rating **5**: Very proactive

**Required Output Format:**
{{
    "rating": int, // a single number on the scale of 1 to 5.
    "justification": str, // explain key interactions or phrases that led to this rating
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}
```

D.4.2 Bot Proactiveness by LLM-as-participant

```
Imagine you are the user in this conversation.
You have just finished talking with an assistant selling phones with offers, accessories and
    additional services.

The conversation you just had:
{conv_history}

Reflecting on your experience as the user in this conversation, please rate the following statement:

"The assistant was proactive addressing my needs with good suggestions and guidance."

Please use the following scale:
- rating **1**: Not at all proactive
- rating **2**: Rather not proactive
- rating **3**: Neither proactive nor not proactive
- rating **4**: Rather proactive
- rating **5**: Very proactive

**Required Output Format:**
{{
    "rating": int, // a single number on the scale of 1 to 5.
    "justification": str, // explain from the user's perspective
                        // what made you feel this way
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}
```

D.5 Goal Achievement

D.5.1 Goal Achievement by LLM-as-observer

```
You are an expert conversation analyst evaluating a real customer interaction
with a sales assistant selling phones, offers, accessories,
```

and additional services.

The conversation you will evaluate involves a real user.
Note that real users often write very short messages
(e.g., "cheap", "yes", "Samsung") this is normal and expected.

The conversation:
{conv_history}

Your goal is to assess **whether the customer achieved their goal** during the conversation.

Please use the following scale:

- rating ****1****: The customer was completely unable to achieve his goal
- rating ****2****: The customer was mostly unable to achieve his goal
- rating ****3****: The customer neither achieved nor failed to achieve their goal
- rating ****4****: The customer mostly achieved his goal
- rating ****5****: The customer fully achieved his goal

****Required Output Format:****

```
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain key interactions or phrases that led to this rating
                        //
                        // IMPORTANT: This is a JSON string value.
                        // - The entire justification must be on one line
                        // - If you need to include a quote within the text, escape it: \"
                        // - Do NOT use single quotes (') for the outer string
                        // - Do NOT use line breaks within the justification
}}
```

D.5.2 Goal Achievement by LLM-as-participant

Imagine you are the user in this conversation.

You have just finished talking with an assistant selling phones with offers, accessories and additional services.

The conversation you just had:
{conv_history}

Reflecting on your experience as the user in this conversation, please rate the following statement:

"I was able to achieve my goal in this interaction."

Please use the following scale:

- rating ****1****: I was completely unable to achieve my goal
- rating ****2****: I was mostly unable to achieve my goal
- rating ****3****: Neither agree nor disagree
- rating ****4****: I mostly achieved my goal
- rating ****5****: I fully achieved my goal

****Required Output Format:****

```
{{
  "rating": int, // a single number on the scale of 1 to 5.
  "justification": str, // explain from the user's perspective
                        // what made you feel this way
}}
```

E LLM Benchmarking Results from Prior Validation Study

Evaluator	User Satisfaction	Bot Helpfulness	Bot Naturalness	Bot Proactiveness	User Goal Achievement
<i>Commercial Models</i>					
GPT-o1	0.76	0.71	0.24	0.71	0.76
GPT-o1-mini	0.73	0.66	0.55	0.62	–
GPT-o3	0.71	0.67	0.51	0.56	0.71
GPT-o3-mini	0.73	0.65	0.50	0.65	0.80
GPT-o4-mini	–	–	–	–	0.76
GPT-4o	0.80	0.76	0.45	0.58	0.78
Claude 3.5 Sonnet v2	0.82	0.73	0.48	0.68	0.73
Claude 3.7 Sonnet	0.71	0.71	0.48	0.69	0.76
<i>Open-Source Models</i>					
Qwen2.5-7B	0.68	0.55	0.28	0.54	0.77
Qwen2.5-14B	0.78	0.59	0.42	0.57	0.75
Qwen2.5-32B	0.68	0.64	0.31	0.63	0.64
Llama-3.1-8B	0.64	0.57	0.52 (*)	0.55	0.76
Selene-1-Mini-Llama-8B	0.65	0.61	0.46	0.65	0.79
DeepSeek-R1-Distill-Qwen-7B	0.55	0.44	0.15	0.38	0.77
DeepSeek-R1-Distill-Qwen-14B	0.72	0.63	0.19	0.49	0.77
DeepSeek-R1-Distill-Llama-8B	0.73	0.52	0.17	0.44	0.77

Table 5: Complete benchmarking results of LLM evaluators against human annotations (n=104 dialogues, except *n=92 for Llama-3.1-8B naturalness) from prior validation studies (Gromada et al., 2025, 2026). All values represent Spearman correlation coefficients (ρ). Bold indicates best performance within each category (commercial/open-source).

F Bot Performance: Baseline vs. Proactive Configuration. Cumulative Link Mixed Models (CLMM) analysis.

F.1 CLMM analysis: per simulator

Dim.	Gemma-3-12b			Gemini-2.0-flash			Gemini-2.5-flash			GPT-4o-mini		
	OR	<i>p</i>	<i>p</i> _{Holm}	OR	<i>p</i>	<i>p</i> _{Holm}	OR	<i>p</i>	<i>p</i> _{Holm}	OR	<i>p</i>	<i>p</i> _{Holm}
User Sat.	0.52	<0.001	0.001	0.55	<0.001	0.004	0.80	0.213	1.00	0.88	0.436	1.00
Bot Help.	0.37	<0.001	<0.001	0.44	<0.001	<0.001	0.74	0.076	0.609	0.44	<0.001	<0.001
Bot Nat.	0.39	<0.001	<0.001	0.40	<0.001	<0.001	0.62	0.032	0.352	0.52	<0.001	0.005
Bot Proact.	0.72	0.047	0.469	0.64	0.005	0.056	0.73	0.058	0.519	0.78	0.106	0.740
Goal Ach.	0.79	0.197	1.00	0.92	0.662	1.00	0.88	0.475	1.00	1.15	0.428	1.00

Table 6: Per-simulator breakdown of odds ratios (OR) and *p*-values for Proactive vs. Base, by dimension and simulator ($n = 312$ dialogs per configuration, per simulator). Raw *p*-values are shown in bold when $p < .05$; Holm-corrected *p*-values (within the 20-test per-simulator family) are shown in parentheses. No Bot Proactiveness or Goal Achievement cell survives Holm correction at the per-simulator level.

F.2 Effect sizes

Dimension	Real Users ($n=104$)	Simulated Users (Pooled, $n=624$)
User Satisfaction	−0.02 [−0.29, 0.25] (negligible)	−0.58 [−0.87, −0.28] (medium)
Bot Helpfulness	−0.01 [−0.28, 0.26] (negligible)	−1.04 [−1.38, −0.70] (large)
Bot Naturalness	−0.27 [−0.54, 0.01] (small)	−1.11 [−1.45, −0.76] (large)
Bot Proactiveness	+0.31 [0.03, 0.59] (small)	−0.54 [−0.83, −0.25] (medium)
Goal Achievement	−0.12 [−0.39, 0.15] (negligible)	−0.17 [−0.44, 0.11] (negligible)

Table 7: Effect sizes (Cohen’s d_z , paired) for Proactive vs. Base, by dimension. Negative values favor Base; positive values favor Proactive. Magnitude labels (negligible/small/medium/large) follow Cohen’s conventions for paired designs.

F.3 TOST Equivalence Tests

Dataset	Dimension	Observed d_z	Eq. bound	TOST p_1	TOST p_2	Verdict
Real Users	User Satisfaction	−0.022	0.2	0.103	0.057	not equivalent
	Bot Helpfulness	−0.012	0.2	0.091	0.006	not equivalent
	Bot Naturalness	−0.265	0.2	0.680	< 0.001	not equivalent
	Bot Proactiveness	0.311	0.2	< 0.001	0.787	not equivalent
	Goal Achievement	−0.114	0.2	0.280	0.013	not equivalent
Simulated Users (pooled)	User Satisfaction	−0.572	0.2	0.996	< 0.001	not equivalent
	Bot Helpfulness	−1.039	0.2	> 0.999	< 0.001	not equivalent
	Bot Naturalness	−1.106	0.2	> 0.999	< 0.001	not equivalent
	Bot Proactiveness	−0.573	0.2	0.991	< 0.001	not equivalent
	Goal Achievement	−0.166	0.2	0.405	0.005	not equivalent

Table 8: TOST equivalence tests for Proactive vs. Base differences, by dataset and dimension. Equivalence bound: $d_z = \pm 0.2$ (small effect). $p_{\text{equiv}} = \max(p_1, p_2)$; $p_{\text{equiv}} < .05$ indicates that the observed effect is statistically equivalent to zero (i.e., falls within the equivalence bounds). None of the comparisons reach equivalence: for real users, this reflects inconclusive null effects (insufficient precision to confirm equivalence); for the pooled simulated-user model, this reflects large effects that exceed the equivalence bounds.

G Ratings distribution on real user dialogues

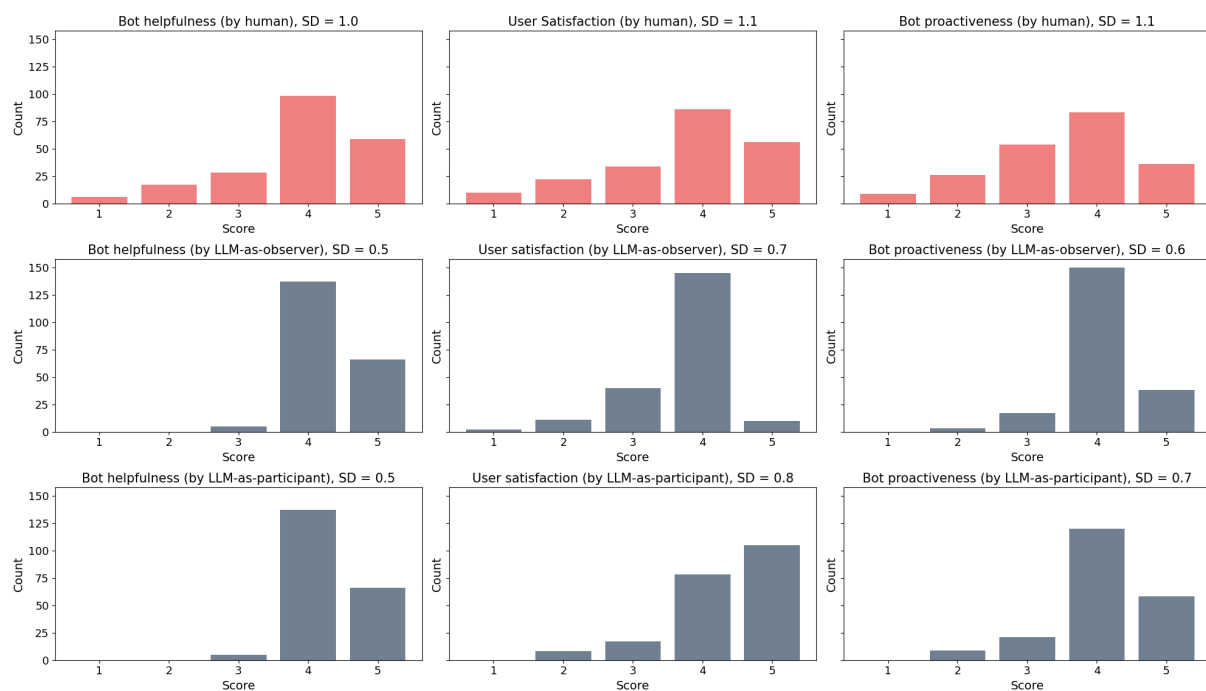


Figure 1: Distribution of human and LLM-as-a-judge ratings (n=208) of dialogues with real users, with standard deviation (SD).