

UXBench: Benchmarking User Experience in AI Assistants

Mengze Hong^{1,2*}, Xia Zeng², Zeyang Lei^{2*}, Sheng Wang², Chen Jason Zhang¹, Di Jiang¹
Taiming Fu², Jinfeng Huang², Mengqiao Liu², Qinghe Chang², Haosheng Zou²
Qiongyi Zhou², Sijun He², Xiaoshuai Chen², Minlong Peng², Di Liang^{2*}

¹Hong Kong Polytechnic University, ²Yuanbao Team, Tencent

Abstract

As AI assistants serve millions of users daily, evaluating user experience (UX) beyond general model capability has become increasingly important. We present **UXBench**, the first user-centric benchmark grounded in real user feedback signals for evaluating preference alignment and dialogue generation. The benchmark consists of three interconnected tasks, UX Judge, UX Eval, and UX Recovery, with 7,400 test instances extracted from over 70K interaction logs of a mainstream Chinese AI assistant. The dataset closely reflects real user distributions, covering 8 scenarios, 83 domains, and diverse failure patterns that pose severe challenges. Extensive experiments on 26 frontier language models provide novel insights into how well models perceive user experience and how improvements in model capability contribute to better dialogue engagement. Through analyses of model behavior and performance gaps, we document six important findings, demonstrating that user feedback prediction is a learnable capability and revealing different aspects that influence user experience. UXBench establishes a new evaluation landscape and calls for greater attention to tailored UX optimization, contributing toward a user-centric scaling law for the development of successful AI assistants. The full project is released at <https://github.com/mengze-hong/UXBench>.

What makes a good product?

Built on functionality, won by experience.

1 Introduction

The long-term success of AI products relies on retaining active users by delivering a positive user experience (UX). Existing LLM benchmarks primarily evaluate foundation model capabilities such as reasoning, knowledge, and instruction following (Mohammadi et al., 2025; Guo et al., 2025), while

largely neglecting user-perceived utility and interaction quality. It remains unclear whether stronger models under existing evaluations necessarily deliver better real-world user experiences, as improvements on traditional benchmarks may translate into only marginal or imperceptible gains for end users who mostly engage in casual information search and chit-chat interactions (Chatterji et al., 2025). This motivates the need for a rigorous UX-centric benchmark grounded in realistic user feedback and interaction behavior to better understand and re-assess progress in AI assistants.

Developing a dedicated benchmark for UX presents three fundamental challenges. First, existing UX research is largely rooted in human-computer interaction (HCI), where experience metrics are often assessed through studies involving human participants that are difficult to scale (Zheng et al., 2025; Zhou et al., 2026b). Second, many evaluation schemes quantify UX through fine-grained dimensions (Mahmud et al., 2025; Li et al., 2026), which, while interpretable, cannot be exhaustively defined or fully operationalized. Intuitively, users do not explicitly deliberate before clicking the “unlike” button; rather, it is an immediate reaction to unsatisfactory responses (Ntoa, 2025; Zheng et al., 2025). Meanwhile, computational approaches such as arena-style benchmarks (Chiang et al., 2024) focus primarily on pairwise model comparisons, which may not align with behavioral signals from real-world interactions, where users typically engage with only one response at a time.

In this paper, we introduce **UXBench**, the first user-centric benchmark designed for automated user experience assessment in AI assistants. Built on large-scale human-AI interaction logs, we propose three interconnected tasks that enable faithful and scalable evaluation of user-perceived utility, revealing key empirical insights for advancing UX benchmarking and model optimization. The main contributions include:

*Corresponding authors.

Benchmark	Source	Scale (N)	Avg Turns	Signal	Judge	UX-driven	User Complains
MT-Bench (Zheng et al., 2023)	Curated	80	2	Rubric	LLM	✗	✗
AlpacaEval (Li et al., 2023)	Synth	805	1	Pairwise	LLM	✗	✗
Chatbot Arena (Chiang et al., 2024)	Logs	Dynamic	1	Voting	Human	✗	✗
Arena-Hard (Li et al., 2025b)	Curated	500	1	Pairwise	LLM	✗	✗
WildBench (Lin et al., 2025)	Logs	1,024	2-3	Rubric	LLM	✗	✗
UXBench (Ours)	Logs	7,400	5.29	Feedback	GRM	✓	✓

Table 1: Comparison of UXBench with existing dialogue-system evaluation benchmarks.

1. **User Experience Modeling:** We leverage user feedback signals from real-time interactions as ground-truth UX labels, turning subjective user experience into an observable and scalable modeling target, and identify three key signal categories that support UX-oriented evaluation and optimization at scale.
2. **Evaluation Framework:** We introduce UXBench, a user-centric benchmark comprising 7,400 test cases sampled from over 70K real interaction logs for evaluating LLMs in user feedback prediction and response generation. A rigorous multi-stage data pipeline enables automated discovery of reliable user feedback signals.
3. **Actionable Insights:** Extensive evaluation on 26 frontier LLMs across three UX tasks reveals six key findings on model performance, failure modes, systematic biases, and scaling trends. A trained reward model further achieves state-of-the-art performance in predicting user feedback, motivating model-based UX optimization.

2 Related Work

2.1 Benchmarking Dialogue Systems

The rise of personalized AI assistants has shifted evaluation beyond traditional QA-based benchmarks toward open-ended dialogue interactions (Yi et al., 2025), in which evaluation protocols primarily follow pairwise, rubric-based, and pointwise paradigms. As shown in Table 1, pairwise and rubric-guided methods are widely used for preference modeling and structured evaluation (Liu et al., 2023; Zhou et al., 2026a), with representative benchmarks such as URS (Wang et al., 2024, 2025b) and WildBench (Lin et al., 2025). More recently, pointwise evaluation has gained attention for improved robustness and interpretability, as shown in Prefix that employs multi-LLM judges

across seven Likert-scale dimensions (Li et al., 2026; Liang et al., 2019a), and is generally more resistant to positional bias and external manipulation (Tripathi et al., 2025). However, existing benchmarks remain limited by poorly aligned LLM judges and insufficient coverage of failure-prone user queries, resulting in an incomplete understanding of user-perceived response quality.

2.2 Benchmarking User Experience

User experience is a central research agenda in HCI, reflecting how effectively, efficiently, and satisfactorily users interact with systems (Ntoa, 2025). Traditional methods such as the User Experience Questionnaire (UEQ) provide structured self-reported feedback but struggle to capture the dynamic, interactive nature of AI systems (Zheng et al., 2025; Lin et al., 2026). Specialized frameworks, such as CASUX (Faruk et al., 2025), have been proposed for dialogue agents to combine qualitative analysis with quantitative UX metrics. However, these approaches remain largely static and depend heavily on human annotators and manually defined evaluation rubrics, limiting their ability to capture the complexity and diversity of real user behavior (Mahmud et al., 2025). UXBench departs from this tradition by leveraging real user feedback signals as direct ground-truth annotations, and by training user-aligned reward models for fully automated evaluation. This reduces reliance on subjective post-hoc judgments and enables more faithful, scalable, and user-aligned model assessment.

2.3 Reward Modeling

Reward models (RMs) are central to modern LLM alignment with user feedback (Ouyang et al., 2022), with an increasing number of studies emphasizing their training and evaluation (Peng et al., 2026; Malik et al., 2026). Early scalar RMs have gradually evolved into generative reward models (GRMs) that produce natural language critiques alongside scalar judgments (Wang et al., 2025a). On the other hand, the quality of RMs has been extensively evaluated

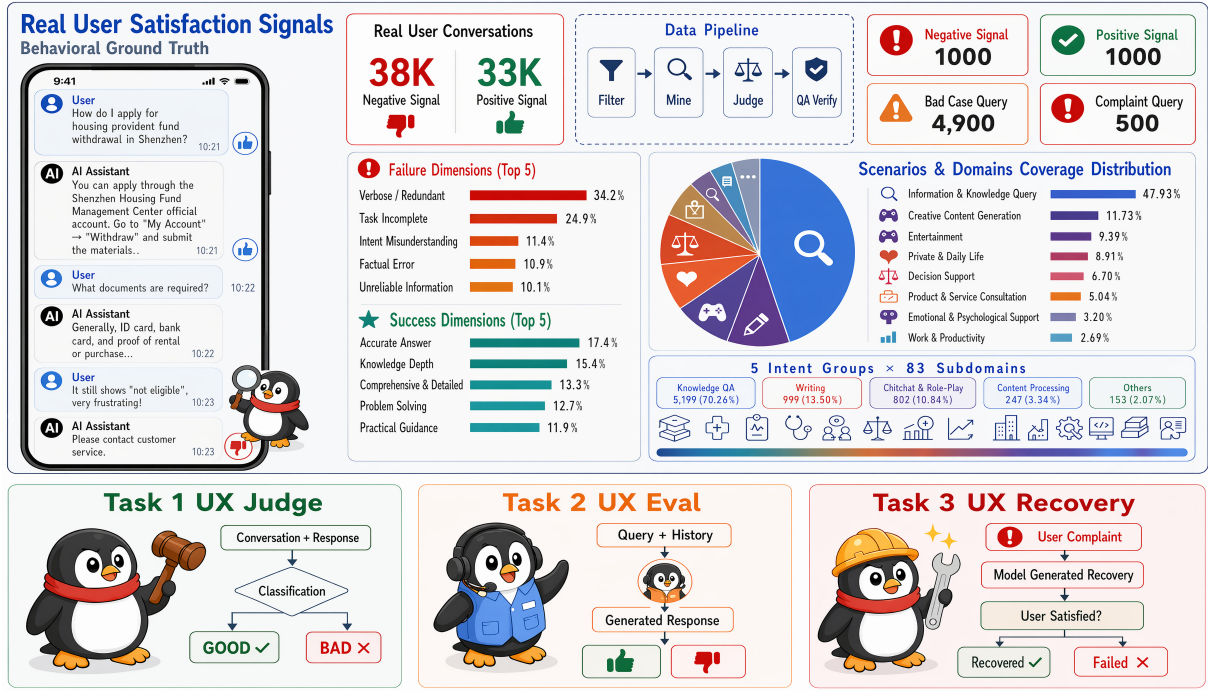


Figure 1: Overview of UXBench, including real-user data sources, data processing pipeline, key statistics across success and failure dimensions, scenario and domain coverage, and an illustration of the three proposed tasks.

through public benchmarks such as RewardBench (Lambert et al., 2025; Xue et al., 2026; Liu et al., 2025a) and PPE (Frick et al., 2025), both of which primarily rely on pairwise ranking against human annotations. In this paper, we leverage RM as a standalone instrument for scalable dialogue evaluation. Unlike conventional LLM-as-a-judge methods (Li et al., 2025a), which rely on prompting or rubric-based evaluation and remain susceptible to biases from imperfect preference alignment (Chen et al., 2024; Liang et al., 2019b; Ye et al., 2025), GRMs trained on authentic user preference signals may provide more human-aligned evaluation.

3 UXBench

3.1 Overview

UXBench is designed to answer three core research questions: (1) how well automated LLM judges predict real user feedback to an AI-generated response; (2) whether frontier LLMs can generate high-quality responses for failure-prone user queries; and (3) how progress in model capability improvements translates into measurable UX gains. Rather than decomposing user experience into isolated dimensions, we formulate positive and negative user feedback as a unified behavioral signal and construct challenging, quality-audited test sets from real interaction logs (Section 3.2). Based

on this formulation, UXBench defines three interconnected tasks (Section 3.3), which progressively assess model capabilities to align with user perception, generate satisfying responses, and recover from service failures. Figure 1 summarizes the data sources, key features, and task design.

To excel in this benchmark, models require not only fundamental reasoning and retrieval capabilities but, more importantly, a user-aligned understanding of what constitutes a good interaction, appropriate dialogue strategies, and sufficient emotional intelligence. Ideally, strong performance on UXBench would directly translate into a better real-world user experience for end users.

3.2 UX Modeling with Feedback Signals

A key bottleneck in user experience evaluation is the lack of reliable ground-truth labels. Existing approaches often rely on human interpretation or post-hoc annotation, making the process subjective and susceptible to annotator bias (Weber-Genzel et al., 2024). To address this challenge, we leverage feedback signals from real-time user interactions and treat them as the most directly observable indicators of response quality (Liu et al., 2025b). Modern AI systems commonly provide explicit feedback mechanisms, such as like and dislike buttons. However, providing such feedback contributes negligibly to task completion and offers

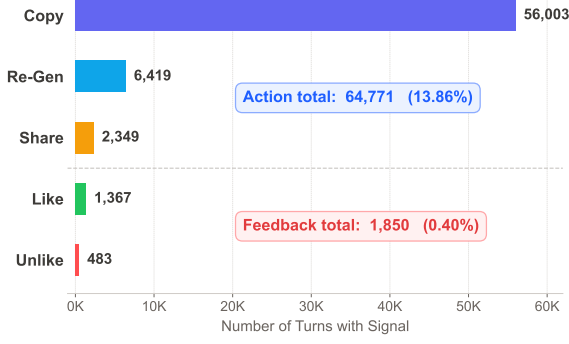


Figure 2: Distribution of explicit feedback signals from 400K randomly sampled dialogue turns.

little immediate benefit to users, making explicit labels reliable but inherently sparse for scalable UX modeling. In a random sample of 400K interaction turns (Figure 2), we find that such feedback signals are extremely sparse compared with action-based behavioral signals.

Action-based signals remain inherently ambiguous and prone to misinterpretation. For example, sharing a response may reflect either satisfaction or an intention to complain. Thus, we complement explicit feedback with **implicit negative feedback** derived from inter-turn user behaviors. Prior work has explored shallow implicit signals, including user input content (Zhang et al., 2018) and statistics such as input length and response latency (Pang et al., 2024). By analyzing turns surrounding explicit dislikes, we derive two rule-based heuristics for implicit negative feedback. First, Skim captures rapid follow-up without fully engaging with the response. Let $\Delta\tau_{t+1} = \tau_{t+1}^u - \tau_t^a$ denote the time to the next user input and $|r_t^a|$ the response length. A turn is flagged when

$$\frac{\Delta\tau_{t+1}}{|r_t^a|} < 0.15 \quad \text{and} \quad \Delta\tau_{t+1} < 30s, \quad (1)$$

with thresholds calibrated on skim cases preceding explicit dislikes. Second, follow-up dissatisfaction captures subsequent Complain/Dismiss turns via fixed dissatisfaction regexes and Repeat/Rephrase turns via high lexical similarity to the preceding query. A locally deployed LLM further verifies heuristic-matched instances for quality assurance. Across 40K sampled dialogue sessions, we identify six negative signals: Repeat (10,337), Rephrase (4,773), Skim (756), Complain (199), Dismiss (187), and explicit Dislike (65; 0.16%), revealing diverse manifestations of user dissatisfaction.

3.3 Task Formulation

Task 1: UX Judge. We first evaluate the model’s ability to predict user feedback from dialogue interactions. Formally, let

$$\mathcal{D} = \{(h_i, q_i, r_i, y_i)\}_{i=1}^N, \quad y_i \in \{y_i^+, y_i^-\} \quad (2)$$

denote a collection of dialogue trajectories, where h_i is the dialogue history, q_i is the user query, r_i is the assistant response, and y_i^+/y_i^- are the corresponding user feedback signals. A language model $\mathcal{M}_\theta$ performs binary UX judgment: $\mathcal{M}_\theta(h_i, q_i, r_i) \mapsto \hat{y}_i \in \{0, 1\}$, predicting latent user satisfaction from interaction content, with accuracy and class-wise recall reported. Although binary prediction is less expressive than rubric-based ordinal ratings, it is motivated by stronger alignment with human judgments under 0–1 loss compared with ordinal or Likert-scale regression settings (Hong et al., 2025b), while directly reflecting real user feedback behavior and providing natural supervision for training and evaluation. In Section 4.1.1, we compare binary judgment with several rubric-based formulations, supporting this design choice as a practical first step toward quantifying user experience.

Task 2: UX Eval. The second task evaluates the model-generated responses to determine whether they lead to a positive user experience. Let $\mathcal{D} = \{(h_i, q_i, y_i^+)\}_{i=1}^N$ denote dialogue sessions extracted exclusively from interactions where the original AI response received explicit negative feedback, forming a dataset intentionally skewed toward failure-prone, hard user cases rather than average interactions with limited discriminative value. Given a generated response $\hat{r}_i$, a reward model $\mathcal{R}_\phi$ predicts a UX score $s_i = \mathcal{R}_\phi(h_i, q_i, \hat{r}_i) \in [0, 1]$, which is converted into a binary preference label $\hat{y}_i = \mathbb{1}[s_i \geq 0.5]$ to report the overall percentage of positive responses:

$$\text{Good\%} = \frac{1}{N} \sum_{i=1}^N \hat{y}_i. \quad (3)$$

Task 3: UX Recovery. The third task extends UX Eval to long-tail, failure-critical scenarios by evaluating whether a model can recover user satisfaction after an explicit interaction failure. Each instance is defined as $d_i = (h_i, c_i)$, where h_i denotes the dialogue history and c_i is a user complaint. From a random sample of 2K complaints,

we observe that over 83% of interactions terminate immediately after a poorly drafted recovery response, highlighting the importance of effective recovery strategies for user retention. We follow the same evaluation protocol as Task 2, while UX Recovery specifically measures whether a model can repair failed interactions through appropriate dialogue strategies, emotional awareness, and corrective behavior under realistic user complaints.

3.4 Dataset Construction Pipeline

UXBench is constructed from 70K+ real user-AI interactions extracted from a mainstream Chinese AI assistant that supports chit-chat, role-play, and information-grounded question answering. Ethical data collection and user privacy are treated as top priorities. All data is collected with explicit user consent through an opt-in experience program, in which a group of users voluntarily authorize the use of their interaction logs for research and service improvement. The pipeline is formulated as a fully automated end-to-end transformation from raw dialogue logs to human-interpretable test cases labeled with feedback signals. A dialogue session $\mathbf{D}_i = \{x_1^u, x_1^a, \dots, x_T^u, x_T^a\}$, first undergoes de-identification and desensitization $\tilde{\mathbf{D}}_i = \phi_{\text{anon}}(\mathbf{D}_i)$ using rule-based and locally deployed LLM (Yang et al., 2025). A deterministic gate ϕ_{gate} then performs data cleaning and removes unsafe cases involving sexual, violent, or other harmful content, after which behavioral feedback signals are derived as $\mathbf{s}_i = \phi_{\text{signal}}(\tilde{\mathbf{D}}_i)$ through pre-defined heuristics.

Given pre-processed dialogue sessions, a multi-agent system constructs test cases using locally deployed Hunyuan and DeepSeek models. A Miner model first proposes failure candidates $b \in \mathcal{C}_i$ covering diverse failure modes, while an independent Judge model assigns quality scores $q_b \in [1, 5]^5$ across five dimensions: completeness, credibility, representativeness, severity, and signal fidelity. Candidates are retained if $\bar{q}_b \geq 4.0$ and $\min(q_b) \geq 3$. Finally, the resulting dataset undergoes LLM-based quality assurance and human expert verification prior to application.

Data Statistics. Table 2 reports key statistics of UXBench across three tasks, comprising 7,400 test cases spanning 8 interaction scenarios and 83 domains. Since the data are randomly sampled from online interaction logs, the scenario distribution closely resembles real-world user interactions, ensuring that the benchmark serves as a reliable proxy

	Task 1 BAD	Task 1 GOOD	Task 2 Eval	Task 3 Recovery
Instances	1,000	1,000	4,900	500
Multi-turn (%)	71.3	71.4	71.7	98.0
Average turns	5.1	5.2	5.2	6.2
Mean query length (chars)	26.1	38.9	31.7	85.9
Mean response length (chars)	836	748	–	–

Table 2: Key statistics of the UXBench dataset.

across diverse user queries. The dataset mainly consists of multi-turn interactions, enabling evaluation under long-context settings. A key feature of UXBench is its structured success and failure taxonomy: negative instances cover 10 failure types, with verbosity being the most common issue, while positive instances span 8 success types, dominated by accurate answering (see Appendix A.3).

4 Results and Discussion

In this section, we progressively analyze LLM capabilities from user feedback prediction to response generation, complemented by key ablation studies for each task. Complete experimental results are reported in Appendix C. Since a key research question is to understand how UX evolves with model capability, we select 26 representative models following two principles: coverage and scale, spanning major model families, capability levels, and release generations.

4.1 Task 1: UX Judgment

While Task 1 is framed as a binary judgment task, Table 3 reveals largely negative results across models. The best-performing zero-shot model (Claude Opus 4.7) achieves 75.3% overall accuracy, with a good accuracy of 89.1% but a substantially lower bad accuracy of 61.5%. This imbalance becomes even more pronounced for weaker models: DeepSeek V3 achieves the highest good accuracy but the lowest bad accuracy at just 11.6%.

Finding 1: Frontier LLMs exhibit a strong and systematic positive bias. Consistent with prior observations that AI models tend to favor AI-generated content (Laurito et al., 2025), all 26 models exhibit a strong bias toward positive judgments and substantially under-detect negative interactions. Since UXBench is class-balanced, a well-calibrated evaluator should achieve comparable performance across both classes. We hypothesize that this failure is driven by the nature of UXBench, where negative samples are primarily grounded in experience-oriented signals such as

Model	Task 1: UX Judge			Task 2: UX Eval	Task 3: UX Recovery
	Acc%	Good-Acc%	Bad-Acc%	Good%	Good%
<i>Google Gemini</i>					
Gemini 2.5 Flash	59.4%	97.5%		36.1%	4.6%
Gemini 2.5 Pro	62.8%	96.8%		50.8%	4.0%
Gemini 3.0 Flash	67.4%	97.7%		52.7%	② 12.7%
Gemini 3.1 Pro	70.4%	91.6%		① 57.1%	9.2%
<i>OpenAI</i>					
GPT-5 mini	65.2%	93.4%		24.0%	3.6%
GPT-5	72.9%	89.5%		34.7%	6.9%
GPT-5.1	72.5%	94.8%		37.1%	7.4%
GPT-5.2	② 75.0%	85.0%	65.1%	30.8%	5.4%
GPT-5.5	③ 74.2%	92.7%		41.2%	9.9%
<i>Anthropic</i>					
Claude Sonnet 4.5	69.3%	89.6%		36.0%	9.0%
Claude Opus 4.5	66.7%	96.7%		37.9%	9.4%
Claude Opus 4.6	72.0%	92.6%		44.3%	① 12.8%
Claude Opus 4.7	① 75.3%	89.1%		44.5%	③ 12.4%
<i>DeepSeek</i>					
DeepSeek R1	58.5%	98.7%		39.5%	5.8%
DeepSeek V3	55.6%	99.7%		35.9%	3.6%
DeepSeek V3.2	64.5%	95.7%		41.2%	7.0%
DeepSeek V4 Pro	64.5%	97.4%		49.7%	11.0%
<i>ByteDance</i>					
Doubao Seed 1.6	57.6%	99.1%		36.8%	6.8%
Doubao Seed 2.0 Lite	57.4%	98.7%		46.3%	10.4%
Doubao Seed 2.0 Pro	60.8%	98.8%		48.7%	10.8%
<i>Others</i>					
Hunyuan 3 Preview	64.3%	95.6%		48.8%	7.6%
GLM-5	66.7%	96.9%		③ 53.0%	10.4%
GLM-5.1	68.5%	96.1%		② 56.6%	11.2%
Kimi K2.5	64.8%	96.8%		50.3%	11.2%
Kimi K2.6	68.7%	96.1%		52.3%	11.4%
Qwen3.6-Plus	65.8%	96.8%		52.3%	12.0%
<i>Trained Model</i>					
Pointwise GRM (ours)	77.2%	82.1%		–	–

Table 3: UXBench results across 26 frontier LLMs and trained Pointwise GRM, grouped by model family.

verbosity, weak engagement, and emotional insensitivity, rather than the factual or lexical errors that LLMs are more familiar with. Despite the strong general capabilities of LLMs, they remain weak at recognizing subtle interaction failures and struggle to accurately predict user feedback, highlighting the difficulty of automated UX evaluation without explicit user-preference alignment.

4.1.1 Comparison of Judgment Formulations

To justify the advantage of unified binary judgment, we first evaluate models under a three-class setting (*positive / neutral / negative*) and examine whether introducing a neutral label can improve UX feedback prediction. As shown in Figure 5, the accuracy consistently drops, with many models assigning difficult BAD cases to the neutral class, forming a clear “neutral escape” pattern. This is

especially evident for Claude models, which map 38.5% of BAD instances to neutral. These results suggest that the observed positive bias is more severe beyond binary settings.

We further compare binary judgment with Likert-style ratings (1–4 to 1–7) and multi-dimensional rubrics from MT-Bench, FLASK, and HelpSteer2. Across Claude Opus 4.7, GPT-5.2, and GPT-5.5, none of these formulations improves alignment with observed user feedback: bad accuracy ranges from 42.6–59.4% under Likert ratings and 20.0–42.4% under rubric scoring. Likert ratings also exhibit only moderate inter-model agreement (Fleiss’ $\kappa = 0.41$ – 0.44). In a human validation study with 30 instances, independent 1–4 scoring achieves only 13.3% agreement with GRM preferences, largely because 60% of pairs receive identical scores, compared with 46.7% agreement for di-

rect pairwise judgment and 83.3% for human annotators. These results do not imply that UX is inherently binary; rather, binary judgment directly matches the observed like/dislike supervision in UXBench while avoiding the additional calibration and aggregation assumptions introduced by finer-grained formulations.

4.1.2 Prompting-based Approach

To understand the effect of explicit prompt instruction, we ablate three prompting strategies for UX judgment across five representative models. Specifically, we evaluate chain-of-thought and few-shot prompting, along with a user modeling scheme that first infers the user profile based on dialogue context before assessing response quality (Figure 6). Figure 7 reports accuracy changes relative to the zero-shot baseline, noting that only user modeling improves performance consistently, with an average gain of 4.54 percentage points. This suggests that UX judgment benefits from modeling the user behind the feedback signal, rather than only matching the surface quality of the response. In contrast, generic reasoning or critique-oriented prompts yield limited and inconsistent gains, suggesting that UX prediction requires user-centered interpretation and personalized modeling efforts.

4.1.3 Training-based Approach

We train a specialized generative reward model (GRM) based on Hunyuan-3 (MoE, ~20B active parameters) as a UX judge, using the same input-output format as the zero-shot evaluator¹. The model is trained for one epoch on 8,547 positive and 8,559 negative in-the-wild instances with explicit user feedback as ground truth. The data undergo standard cleaning and de-identification without LLM-based filtering or manual curation, enabling an automated data flywheel. To prevent leakage, training data are sampled from May 2026, temporally disjoint from the UXBench test set collected from January to March 2026. We use a learning rate of 4×10^{-6} , a global batch size of 128, and 128 NVIDIA H20 GPUs. The training loss rapidly converges and stabilizes, demonstrating the feasibility of efficiently training UX judges directly from in-the-wild user feedback.

Finding 2: Pointwise GRM aligns strongly with real user feedback. As shown in Table 3, the trained GRM is well calibrated and outperforms

all frontier models in both overall accuracy and bad-case recall. Considering its relatively small parameter size, the trained GRM serves as both an effective and efficient evaluation tool, enabling lightweight deployment in real-time service systems. These promising results suggest that **UX judgment is a learnable capability** grounded in real user feedback signals, rather than merely an emergent property of large-scale pretraining. The unsaturated scores further position UXBench as a useful benchmark for evaluating future GRMs, enabling continuous improvement in user alignment.

4.2 Task 2: UX Evaluation

Task 2 examines which model produces the most satisfying responses to users, with all outputs evaluated by the trained GRM under the same setting. Table 3 shows that the best-performing model achieves only a 57.1% good response rate, highlighting the difficulty of UXBench and leaving substantial room for improvement. Interestingly, Claude Opus 4.7, a state-of-the-art coding model, underperforms more general models such as DeepSeek V4 Pro, suggesting that user-perceived utility does not necessarily depend on performance on complex tasks, since user queries are mostly focused on common QA and chit-chat.

To analyze performance gaps between models, we manually inspect cases where responses to the same query receive opposite UX judgments. A key factor is conversational style, particularly the opening sentence. For example, “你说得对，我之前说的不准确” (“You are right. What I said earlier was inaccurate”) is more direct and conversational than the more formulaic “感谢您的指正，我之前的回答存在错误” (“Thank you for pointing that out. My previous response contained an error”). Across cases, positive UX is associated with natural and direct openings (Okoso et al., 2025), accurate inference of implicit user needs, appropriate use of conversational context without fabricating prior information, and practical guidance rather than excessive conservatism. We also observe advantages of Chinese LLMs in information- and knowledge-intensive scenarios requiring tailored knowledge, consistent with prior findings (Hong et al., 2025a).

Finding 3: User experience exhibits a positive but weak scaling trend with model evolution.

As shown in Figure 3, improvements in model capability generally translate into better user experience, with newer generations generally outperform-

¹<https://huggingface.co/mengze-hong/UXBench-GRM>

Label	Share	Definition	Real Response Opening Example
Apology	45.98%	Apologizes, admits fault, or takes responsibility.	抱歉刚才太多废话了，直接给答案... <i>Sorry, I was too verbose just now; here is the direct answer...</i>
Agreement	28.64%	Validates the complaint or agrees that the user is right.	你说得对，我停下来认真反思一下。 <i>You are right; let me stop and reflect carefully.</i>
Error Diagnosis	14.67%	Explains the error or states an information boundary.	刚才我确实理解跑偏了，把你的需求想得复杂了。 <i>I did misunderstand earlier and made your request more complicated than it was.</i>
Humor	4.81%	Uses humor or de-escalation to repair rapport.	哥们儿消消气，是我反复横跳把你绕晕了，这锅我背。 <i>Buddy, take it easy; I kept switching directions and confused you. That's on me.</i>
Direct Fix	3.63%	Directly provides the correction or the next step.	11岁女孩正常体重约31–44千克，中位数约37千克。 <i>For an 11-year-old girl, normal weight is about 31–44 kg, median around 37 kg.</i>
Clarification	1.56%	Asks what was wrong or what information is needed.	是我哪里没帮上忙吗？你直接告诉我哪里不对。 <i>Did I fail to help somewhere? Tell me directly what was wrong.</i>

Table 4: Failure recovery strategies with real examples.

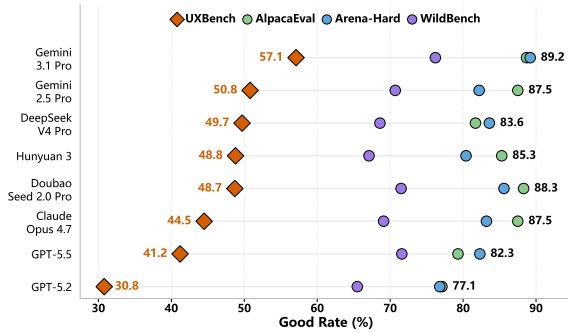


Figure 4: Cross-benchmark generalization.

Finding 6: Response strategies contribute directly to UX recovery. We identify six recovery strategies (Table 4), with models relying predominantly on social repair rather than direct task correction: Apology accounts for 45.98% of responses, while Direct Fix accounts for only 3.63%. Recovery strategies also vary substantially across model families (Figure 11). Poorer-performing models, such as the GPT family, tend to overemphasize error diagnosis, explaining failures rather than directly addressing user needs, whereas stronger models, such as Claude, more often acknowledge complaints and use humor or playful tones to defuse tension. DeepSeek-V3, the worst-performing model, relies most heavily on clarification, suggesting that users who explicitly complain generally benefit more from immediate correction than further requests for information.

4.4 Cross-Benchmark Generalization

Finally, we validate our findings on three public English dialogue generation benchmarks under identical inference settings, using the trained GRM for evaluation. As shown in Figure 4, models achieve much higher average scores on Arena-Hard, WildBench, and AlpacaEval than on UXBench, with the average dropping from 84.1%, 81.8%, and 69.1%

to only 45.9%, indicating the pronounced challenges posed by failure-prone user queries. Rankings are broadly consistent across benchmarks, with several shifts highlighting the UX-specific focus: strong models on English benchmarks, such as GPT-5.5, rank lower on UXBench, while DeepSeek V4 Pro becomes more competitive. This suggests that UXBench evaluates AI assistants beyond general instruction-following, positioning the proposed benchmark as a valuable resource for assessing user-experience-oriented model behavior in real-world applications.

5 Conclusion

This paper presents **UXBench**, the first benchmark grounded in real user feedback signals for evaluating AI assistants beyond capability-oriented proxies. Built from large-scale interaction logs, UXBench defines three tasks covering user feedback prediction, response generation, and failure recovery, and reveals consistent gaps across 26 frontier LLMs, including systematic positive bias, self-preference, and a weak-scaling trend that indicates a potential mismatch between model capability and user-perceived utility. UXBench advances user-centric model evaluation by emphasizing interaction quality, motivating future model development toward better user preference alignment and more robust response strategies. Furthermore, UXBench highlights the potential of leveraging real user feedback signals to alleviate the shortage of explainable ground-truth annotations, enabling fully automated evaluation at scale. Finally, we call for future human-subject research to better understand the evolving expectations of AI assistants and how they align with model evolution, with the goal of discovering the user-centric scaling laws that can guide future model advancements toward the long-term success of commercial AI products.

Limitations

UXBench has several limitations. First, it is primarily grounded in Chinese-language interactions and Chinese-specific domains. Although results on three English benchmarks show broadly consistent patterns, extending UXBench to multilingual and cross-cultural settings would improve its generalizability. Second, the response generation task evaluates single-turn generation conditioned on multi-turn history rather than continuous interaction, where user experience may evolve over time.

Finally, UXBench models user feedback through binary judgments aligned with observable like/dislike signals. While practical for predicting real user feedback, this formulation does not fully capture the multi-dimensional nature of UX or explain why users are satisfied or dissatisfied. Future work could incorporate fine-grained diagnostic dimensions, personalized user context, and interactive evaluation protocols to enrich the discussion and provide a more detailed analysis of user experience in modern AI assistants.

Ethical Considerations

Ethical data collection and user privacy are treated as top priorities. All data is collected under explicit user consent through an opt-in experience program, where a group of users voluntarily authorize the use of their interaction logs for research and service improvement. Before release, all data underwent strict de-identification to prevent disclosure of private information. The released data should be used under privacy-preserving protocols and only for model evaluation, not for model training.

Acknowledgements

We express our gratitude to the Tencent Yuanbao team and gratefully acknowledge Simonjmdeng, Haojing Huang, Zijian Li, Lucas Mu Li, Fubao Zhang, Mona Zhou, Wei Ma, Yuan Hua, Qi Zhu, Shuo Jiang, Chenxuan Ma, Yuanmeng Zhang, Jian Song, and Davey Chen for their valuable contributions, guidance, and support throughout this work.

References

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use chatgpt](#). Working Paper 34255, National Bureau of Economic Research.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *EMNLP 2024*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating LLMs by human preference](#). In *ICLR*. PMLR.

Lawal Ibrahim Dutsinma Faruk, Debajyoti Pal, Suree Funilkul, Thinagaran Perumal, and Pornchai Mongkolnam. 2025. [Introducing casux: A standardized scale for measuring the user experience of artificial intelligence based conversational agents](#). *International Journal of Human-Computer Interaction*, 41(9):5274–5298.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2025. [How to evaluate reward models for rlhf](#). In *ICLR*, volume 2025, pages 18128–18163.

Cheng Guo, Hu Kai, Shuxian Liang, Yiyang Jiang, Yi Gao, Xian-Sheng Hua, and Wei Dong. 2025. [SD-Bench: A survey-based domain-specific LLM benchmarking and optimization framework](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13492–13506, Vienna, Austria. Association for Computational Linguistics.

Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. 2025a. [QualBench: Benchmarking Chinese LLMs with localized professional qualifications for vertical domain evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5938–5953, Suzhou, China. Association for Computational Linguistics.

Mengze Hong, Wailing Ng, Chen Jason Zhang, Yuanfeng Song, and Di Jiang. 2025b. [Dial-in LLM: Human-aligned LLM-in-the-loop intent clustering for customer service dialogues](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5885–5900, Suzhou, China. Association for Computational Linguistics.

Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee, and Jaegul Choo. 2025. [The comparative trap: Pairwise comparisons amplifies biased preferences of LLM evaluators](#). In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 79–108, Suzhou, China. Association for Computational Linguistics.

Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. [Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. **RewardBench: Evaluating reward models for language modeling**. In *NAACL 2025*.
- Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. 2025. **Ai-ai bias: Large language models favor communications generated by large language models**. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. **From generation to judgment: Opportunities and challenges of LLM-as-a-judge**. In *EMNLP 2025*.
- Jialin Li, Zhenhao Chen, Hanjun Luo, and Hanan Salam. 2026. **Prefix: Understand and adapt to user preference in human-agent interaction**. In *ACL 2026*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025b. **From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline**. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34209–34231. PMLR.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. **AlpacaEval: An automatic evaluator of instruction-following models**. https://github.com/tatsu-lab/alpaca_eval.
- Di Liang, Fubao Zhang, Qi Zhang, and Xuan-Jing Huang. 2019a. Asynchronous deep interaction network for natural language inference. 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 2692–2700.
- Di Liang, Fubao Zhang, Weidong Zhang, Qi Zhang, Jinlan Fu, Minlong Peng, Tao Gui, and Xuanjing Huang. 2019b. Adaptive multi-attention network incorporating answer information for duplicate question detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 95–104.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. **Wildbench: Benchmarking llms with challenging tasks from real users in the wild**. In *International Conference on Learning Representations*, volume 2025, pages 47852–47870.
- Zekai Lin, Chao Xue, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Lei Jiang, Yu Lu, Bob Simons, Shuang Liang, and Minlong Peng. 2026. **Parameter importance is not static: Evolving parameter isolation for supervised fine-tuning**. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32846–32864, San Diego, California, United States. Association for Computational Linguistics.
- Xiaoyu Liu, Di Liang, Hongyu Shan, Peiyang Liu, Yonghao Liu, Muling Wu, Yuntao Li, Xianjie Wu, Li Miao, Jiangrong Shen, and Minlong Peng. 2025a. **Structural reward model: Enhancing interpretability, efficiency, and scalability in reward modeling**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 672–685, Suzhou (China). Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. **G-eval: NLG evaluation using gpt-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. 2025b. **User feedback in human-LLM dialogues: A lens to understand users but noisy as a learning signal**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2666–2681, Suzhou, China. Association for Computational Linguistics.
- Raj Mahmud, Yufeng Wu, Abdullah Bin Sawad, Shlomo Berkovsky, Mukesh Prasad, and A. Baki Kocaballi. 2025. **Evaluating user experience in conversational recommender systems: A systematic review across classical and llm-powered approaches**. In *Proceedings of the 37th Australian Conference on Human-Computer Interaction, OzCHI '25*, page 81–93, New York, NY, USA. Association for Computing Machinery.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. 2026. **Rewardbench 2: Advancing reward model evaluation**. In *International Conference on Learning Representations*, volume 2026, pages 144839–144866.
- Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. **Evaluation and benchmarking of llm agents: A survey**. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25*, page 6129–6139, New York, NY, USA. Association for Computing Machinery.
- Stavroula Ntoa. 2025. **Usability and user experience evaluation in intelligent environments: A review and reappraisal**. *International Journal of Human-Computer Interaction*, 41(5):2829–2858.

- Ayano Okoso, Mingzhe Yang, and Yukino Baba. 2025. [Do expressions change decisions? exploring the impact of ai’s explanation tone on decision-making](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Richard Yuanzhe Pang, Stephen Roller, Kyunghyun Cho, He He, and Jason Weston. 2024. [Leveraging implicit feedback from deployment data in dialogue](#). In *EACL*, pages 60–75, St. Julian’s, Malta. Association for Computational Linguistics.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Lei Hou, and Juanzi Li. 2026. [WildReward: Learning reward models from in-the-wild human interactions](#). In *ACL*.
- Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. [Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation](#). In *Second Conference on Language Modeling*.
- Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Qiaozhi He, Murun Yang, Bei Li, Tong Xiao, Chunliang Zhang, Tongran Liu, and Jingbo Zhu. 2025a. [GRAM: A generative foundation reward model for reward generalization](#). In *ICML*. PMLR.
- Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. [A user-centric multi-intent benchmark for evaluating large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.
- Yao Wang, Di Liang, and Minlong Peng. 2025b. Not all parameters are created equal: Smart isolation boosts fine-tuning performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9896.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating annotation error from human label variation](#). In *ACL*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.
- Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. 2026. [Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30186–30213, San Diego, California, United States. Association for Computational Linguistics.
- Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. [Robust utility-preserving text anonymization based on large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria. Association for Computational Linguistics.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. 2025. [Justice or prejudice? quantifying biases in llm-as-a-judge](#). In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.
- Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. [A survey on recent advances in llm-based multi-turn dialogue systems](#). *ACM Comput. Surv.*, 58(6).
- Wei-Nan Zhang, Lingzhi Li, Dongyan Cao, and Ting Liu. 2018. [Exploring implicit feedback for open domain conversation generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Qingxiao Zheng, Minrui Chen, Pranav Sharma, Yiliu Tang, Mehul Oswal, Yiren Liu, and Yun Huang. 2025. [Evalignux: Advancing ux evaluation through llm-supported metrics exploration](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. 2026a. [RubricBench: Aligning model-generated rubrics with human standards](#). In *ACL 2026*, pages 31179–31200, San Diego, California, United States. Association for Computational Linguistics.
- Zihan Zhou, Yanan Liu, Yuyang Xie, Bin Wang, Xiaochun Yang, and Zezheng Feng. 2026b. [Diaglink: A dual-user diagnostic assistance system by synergizing experts with llms and knowledge graphs](#). In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI ’26, New York, NY, USA. Association for Computing Machinery.

Quality Dimension	BAD	GOOD
Overall quality = high	98.5%	99.8%
Representativeness = high	93.5%	99.4%
Signal confidence = high	97.2%	99.7%
Judge avg score (1-5)	4.62	4.63

Table 5: Data quality statistics for UXBench.

A Details of UXBench Dataset

A.1 Data Quality and Human Validation

We validate UXBench data quality using both LLM-based quality assurance and blind human annotation. Each instance is assessed by an independent LLM judge along multiple dimensions, including overall quality, representativeness, and signal confidence. As shown in Table 5, both BAD and GOOD instances in Task 1 achieve consistently high scores, with 98.5% of BAD cases and 99.8% of GOOD cases rated as high-quality.

We further recruit three human experts to independently perform blind GOOD/BAD judgments on Task 1 instances, testing the alignment between the extracted feedback signals (ground truth) and human perception. Each validation group contains 50 cases, with 25 GOOD and 25 BAD instances. The experts achieve overall accuracies of 68.0%, 74.0%, and 66.0%, respectively, with BAD recognition accuracy ranging from 72.0% to 76.0%. These results indicate that UXBench labels are meaningfully recoverable by human judgment while confirming that the task remains challenging even for humans, as many cases require reasoning over subtle dissatisfaction cues, multi-turn context, and whether the response fully satisfies the user’s underlying intent.

A.2 Human Annotator Profiles

Each annotator has over three years of professional experience in customer service call centers in mainland China. All annotators are fluent in Chinese and experienced in dialogue quality assessment. Unlike task-specific annotation settings, no additional training examples or calibration sessions were provided before annotation. Annotators were given only the task instruction and asked to independently judge whether each response would satisfy the user under the pointwise or pairwise convention. The annotation was conducted in a single continuous session to better reflect natural human judgment under a blind evaluation setting, without alternating preferences or post-hoc corrections.

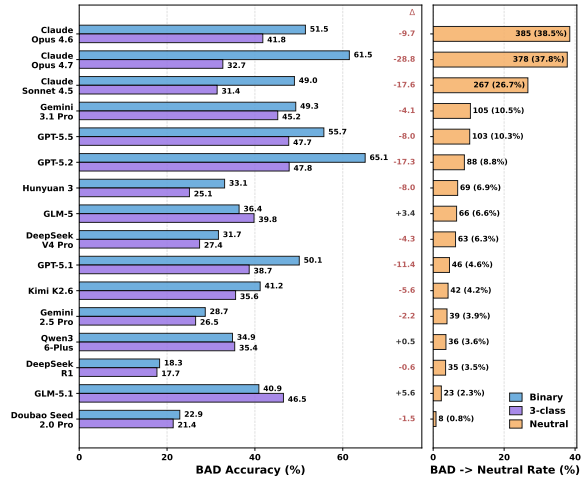


Figure 5: Comparison of binary and three-class judgment formulation.

A.3 Failure and Success Dimensions

We analyze the causes of user satisfaction and dissatisfaction and present the resulting taxonomy in Table 6. Specifically, we ask Claude Opus 4.7 to label each instance with a free-text failure or success dimension, followed by clustering and human verification. Negative-feedback instances span 10 failure dimensions, dominated by recurring interaction failures such as verbosity/redundancy, task incompleteness, and intent misunderstanding, suggesting that dissatisfaction often arises from interaction friction rather than a single catastrophic mistake. In contrast, positive-feedback instances span 8 success dimensions and are primarily associated with accurate answers, knowledge depth, comprehensiveness, and effective problem-solving. Overall, positive dimensions are more capability-oriented, whereas negative dimensions are frequently interaction-oriented, highlighting that strong user experience depends on both model capability and alignment with user expectations.

B Model Selection and Inference

All models are accessed through a LiteLLM-compatible Chat Completions API using a unified request format across model families. For reasoning-capable models, we enable the strongest available reasoning mode with `reasoning_effort=high`, while models without explicit reasoning controls use their default inference settings. For judgment tasks, model outputs are parsed into binary verdicts, while for generation tasks, we retain the final assistant response and the API-returned reasoning content.

BAD: Failure Dimensions (N=1,000)			GOOD: Success Dimensions (N=1,000)		
Dimension	Count	%	Dimension	Count	%
Verbosity / Redundancy	343	34.3	Accurate Answering	174	17.4
Task Incompleteness	248	24.8	Knowledge Depth	154	15.4
Intent Misunderstanding	115	11.5	Comprehensive Detail	137	13.7
Factual Error	110	11.0	Problem Solving	126	12.6
Information Reliability Issue	101	10.1	Practical Guidance / Actionability	121	12.1
Instruction / Format Failure	28	2.8	Creative Generation	116	11.6
Emotional Tone Mismatch	20	2.0	Task Completion	90	9.0
Insufficiently Informative	22	2.2	Empathetic Support	82	8.2
Safety / Refusal Issue	9	0.9			
System / Technical Error	4	0.4			

Table 6: Taxonomy of failure and success dimensions in Task 1 user feedback prediction.

Family	Models
Gemini	2.5 Flash, 2.5 Pro, 3.0 Flash, 3.1 Pro
GPT	5 mini, 5, 5.1, 5.2, 5.5
Claude	Sonnet 4.5, Opus 4.5, 4.6, 4.7
DeepSeek	R1, V3, V3.2, V4 Pro
Doubao Seed	1.6, 2.0 Lite, 2.0 Pro
Others	Kimi K2.5, K2.6; GLM-5, 5.1
	Qwen3.6-Plus, Hunyuan 3

Table 7: Model selection for UXBench evaluation.

C Full Experimental Results

C.1 Analysis of Bias Behavior

Table 8 shows that pointwise LLM-as-a-judge evaluation produces highly unstable scores across judges. Gemini-based judges assign extremely high Good rates to almost all model outputs, often above 90%, substantially weakening leaderboard separability, while GPT-5.2 produces much lower scores for most non-GPT models and ranks GPT-family models at the top. Table 9 further shows that this bias persists in pairwise comparison: GPT-5.2 strongly favors GPT-family responses, assigning GPT-5.1 or GPT-5.5 win rates of 75.8%–83.2% against strong non-GPT models, while Gemini 3.1 Pro similarly prefers Gemini 3.1 Pro over Claude Opus 4.7, DeepSeek V4 Pro, and GPT-5.1 with win rates above 74%. In contrast, Claude Opus 4.7 consistently under-prefers Claude-family responses. Together, these results indicate that LLM judges are not family-neutral for UX evaluation, exhibiting systematic self-preference or judge-specific calibration bias that substantially distorts model rankings.

C.2 Evaluation with Open-Weight Small Language Models

We evaluate two open-weight small language models. As shown in Table 10, their UX-driven capabilities remain limited. Qwen3-8B consistently

Model	Good % (under different judges)		
	GPT-5.2	Gemini 2.5 Flash	Gemini 3.1 Pro
GPT-5.2	78.3%	95.5%	84.5%
GPT-5.1	70.0%	96.1%	85.3%
GPT-5.5	68.8%	93.9%	90.8%
Gemini 2.5 Flash	45.3%	94.4%	69.5%
Gemini 2.5 Pro	43.7%	96.1%	80.6%
Gemini 3.1 Pro	41.5%	97.3%	95.9%
Gemini 3.0 Pro	41.3%	95.4%	90.3%
Gemini 3.0 Flash	40.7%	94.9%	85.5%
Claude Opus 4.7	44.3%	93.8%	82.3%
Claude Sonnet 4.5	40.4%	91.5%	74.2%
DeepSeek V4 Pro	42.8%	94.4%	83.6%
DeepSeek R1	33.5%	93.3%	71.0%
Doubao Seed 2.0 Pro	43.7%	95.8%	85.5%
Kimi K2.6	54.3%	94.5%	90.5%
Qwen3.6-Plus	52.6%	96.1%	87.7%
GLM-5.1	40.5%	96.3%	89.1%
GLM-5	40.3%	95.5%	87.5%
Hunyuan 3 Preview	37.7%	95.1%	78.3%

Table 8: Evaluation of model-generated responses with different judge models.

outperforms Qwen2.5-7B in overall Task 1 accuracy and achieves higher GRM-rated quality in both Task 2 and Task 3, suggesting that newer models provide stronger UX capabilities, consistent with the scaling pattern observed in the main results. However, the low accuracy in Task 1 and low Good rates in generation and recovery tasks indicate that small local models still require task-specific optimization before they can reliably capture dissatisfaction, generate satisfying responses, or recover from service failures in dialogue interactions.

C.3 Impact of Prompting Strategies

Figure 7 reports the impact of different prompting strategies on UX Judge. Among the three strategies, only the **user modeling prompt** improves performance across four models without degrading the fifth, yielding an average gain of 4.54 percentage points. By explicitly inferring the user’s likely profile before judgment (Figure 6), user modeling provides more consistent gains than generic rea-

Judge	Model A	Model B	A Win%	B Win%	Tie%
<i>GPT-5.2 as judge</i>					
	GPT-5.1	Claude Opus 4.7	83.2%	16.6%	0.2%
	GPT-5.1	DeepSeek V4 Pro	81.6%	17.8%	0.6%
	GPT-5.1	Gemini 3.1 Pro	80.6%	19.2%	0.2%
	GPT-5.5	Claude Opus 4.7	75.8%	23.4%	0.8%
	GPT-5.5	Gemini 3.1 Pro	79.0%	20.4%	0.6%
<i>Gemini 3.1 Pro as judge</i>					
	Gemini 3.1 Pro	Claude Opus 4.7	81.2%	15.0%	3.8%
	Gemini 3.1 Pro	DeepSeek V4 Pro	74.4%	20.0%	5.6%
	Gemini 3.1 Pro	GPT-5.1	83.8%	13.2%	3.0%
	Gemini 2.5 Pro	Claude Opus 4.7	46.8%	50.2%	3.0%
	Gemini 2.5 Pro	GPT-5.1	50.6%	47.4%	2.0%
<i>Claude Opus 4.7 as judge</i>					
	Claude Opus 4.7	DeepSeek V4 Pro	39.6%	58.6%	1.8%
	Claude Opus 4.7	Gemini 3.1 Pro	24.4%	74.2%	1.4%
	Claude Opus 4.7	GPT-5.1	41.2%	57.8%	1.0%
	Claude Sonnet 4.5	Gemini 3.1 Pro	10.6%	88.4%	1.0%
	Claude Sonnet 4.5	GPT-5.1	20.0%	79.2%	0.8%

Table 9: Pairwise evaluation across three judge models on 500 randomly sampled test cases.

Task	Metric	Qwen3-8B	Qwen2.5-7B
Task 1	Overall Acc.	58.98	55.66
	GOOD Acc.	97.80	83.58
	BAD Acc.	20.12	26.73
Task 2	Good (%)	23.20	9.40
Task 3	Good (%)	4.20	1.80

Table 10: Performance of open-weight small LMs.

User Modeling Prompt for UX Judge

Given the dialogue history, user query, and assistant response, first infer the user’s likely profile before making the UX judgment.

Specifically, consider: (1) cognitive level and domain familiarity; (2) interaction goal; (3) emotional state; (4) expected response style and information need.

Then decide whether the assistant’s response would satisfy this user under the inferred profile.

Output only a JSON object:

{"verdict": 1} if the user is likely satisfied.

{"verdict": -1} if the user is likely dissatisfied.

Figure 6: Prompt template for the user modeling.

soning or critique-oriented prompting, suggesting that UX prediction benefits from user-centered interpretation. For UX Eval, we compare four system prompts using Gemini 3.1 Pro Preview on 1,000 randomly sampled test cases. As shown in Figure 9, different system prompts yield negligible performance differences, suggesting that simple prompt steering provides limited benefit on diverse, failure-prone user queries.

For UX Recovery, Figure 8 shows that replacing the recovery prompt with a generic helpful-assistant prompt reduces performance, highlighting the importance of failure-aware instructions. CoT prompting consistently improves performance across all four models, whereas failure-aware and critique-based prompts provide only marginal or inconsistent gains. Nevertheless, the best performance remains below 18%, indicating that prompting alone is insufficient to substantially improve failure recovery in the presence of explicit user complaints. Overall, prompting methods appear to be less effective in UX-centric optimization.

C.4 Impact of Reasoning Effort

We study whether increasing reasoning effort improves model performance by comparing high- and low-reasoning settings across representative models. As shown in Figure 10, higher reasoning effort provides the clearest benefit in Task 1, with Claude Opus 4.7 improving from 50.3% to 61.5% BAD accuracy and Gemini 3.1 Pro from 42.1% to 49.3%, suggesting that additional reasoning helps identify subtle signals of dissatisfaction. In contrast, the gains are small and inconsistent in Task 2 and marginal in Task 3. These results suggest that response generation and failure recovery depend less on reasoning depth alone and more on response style, user alignment, and the ability to produce an appropriate response strategy.

C.5 Cross-Benchmark Generalization

In support of the discussion in Section 4.4, we compare UXBench with three widely used English benchmarks. As shown in Table 11, models achieve much higher average scores on Arena-Hard, Wild-

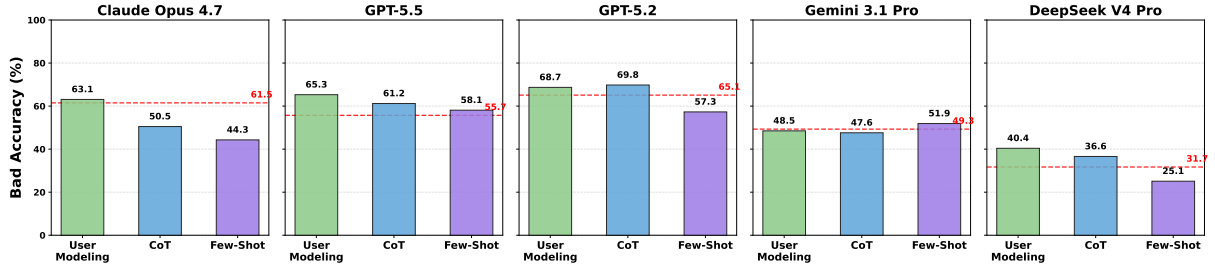


Figure 7: Comparison of negative feedback prediction accuracy (task 1) across different system prompt variants, with the helpful-assistant prompt (red dashed line) shown as baseline.

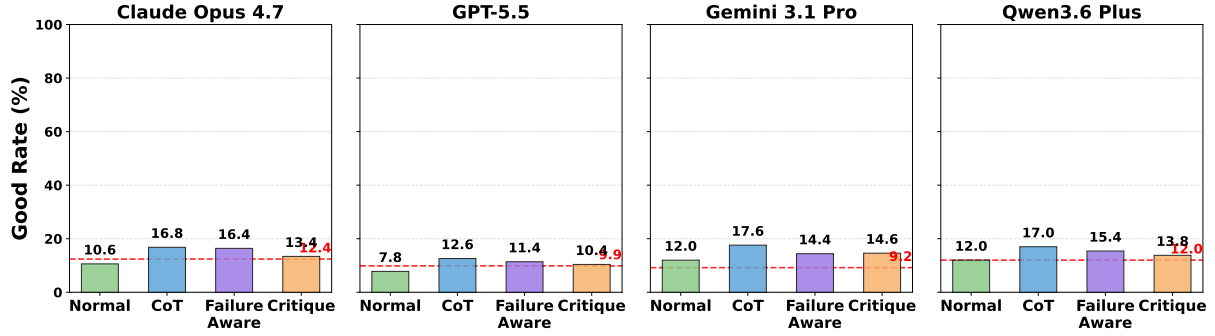


Figure 8: Comparison of failure recovery performance (task 3) across system prompt variants in Task 3, with the helpful-assistant prompt shown as the baseline.

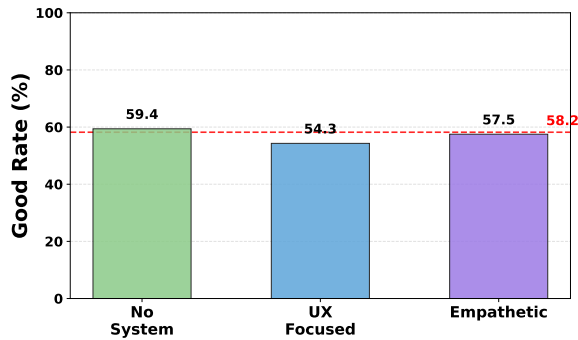


Figure 9: Comparison of response generation (task 2) with Gemini 3.1 Pro and different system prompts.

mance varies substantially across scenarios and tasks. Across models, Task 2 consistently performs strongly on *Daily* scenarios, whereas Task 3 remains substantially more challenging across nearly all scenarios. Performance also varies across interaction contexts, with emotionally sensitive scenarios remaining particularly challenging for several models. Overall, the scenario-level results demonstrate that UXBench difficulty is not uniform across domains and that model strengths do not necessarily generalize across different user scenarios or interaction stages.

Bench, and AlpacaEval than on UXBench, with average scores decreasing from 84.1%, 81.8%, and 69.1%, respectively, to 45.9% on UXBench, highlighting the challenge posed by failure-prone user queries. Model rankings remain broadly consistent across benchmarks, suggesting that UXBench still reflects general response quality. However, several ranking shifts highlight its UX-specific focus, suggesting that UXBench captures capabilities beyond general instruction following, including sensitivity to user dissatisfaction and service-quality failures.

C.6 Performance Breakdown by Scenarios

Finally, we break down model performance by user scenario. As shown in Figure 12, perfor-

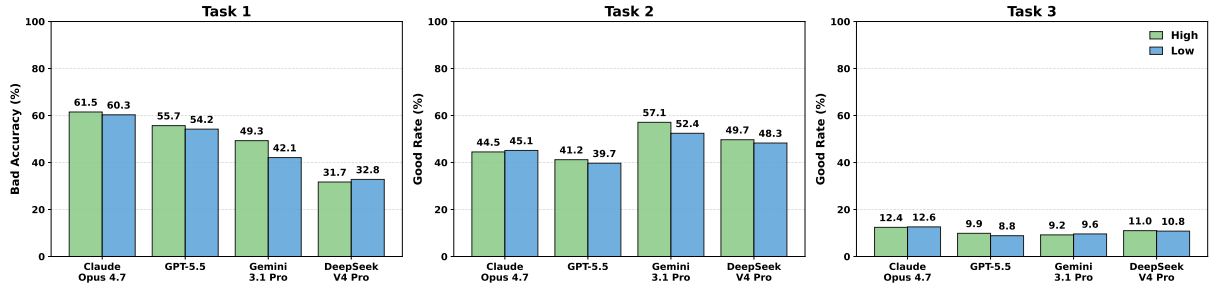


Figure 10: Effect of reasoning effort on performance across the three UXBench tasks, comparing high- and low-reasoning settings for representative models.

Model	Family	UXBench (N=4,900)	Arena-Hard (N=500)	WildBench (N=1,024)	AlpacaEval (N=805)
Gemini 3.1 Pro	Gemini	57.1% (1)	88.7% (1)	89.2% (1)	76.2% (1)
Gemini 2.5 Pro	Gemini	50.8% (2)	87.5% (3)	82.2% (6)	70.7% (4)
DeepSeek V4 Pro	DeepSeek	49.7% (3)	81.7% (6)	83.6% (3)	68.6% (6)
Hunyuan 3 Preview	Hunyuan	48.8% (4)	85.3% (5)	80.4% (7)	67.1% (7)
Doubao Seed 2.0 Pro	Doubao	48.7% (5)	88.3% (2)	85.6% (2)	71.5% (2)
Claude Opus 4.7	Claude	44.5% (6)	87.5% (3)	83.2% (4)	69.1% (5)
DeepSeek V3.2	DeepSeek	41.2% (7)	81.7% (6)	73.2% (9)	61.6% (9)
GPT-5.5	OpenAI	41.2% (7)	79.3% (8)	82.3% (5)	71.6% (2)
GPT-5.2	OpenAI	30.8% (9)	77.1% (9)	76.8% (8)	65.5% (8)
Average		45.9%	84.1%	81.8%	69.1%

Table 11: Performance comparison with three English benchmarks on response generation.

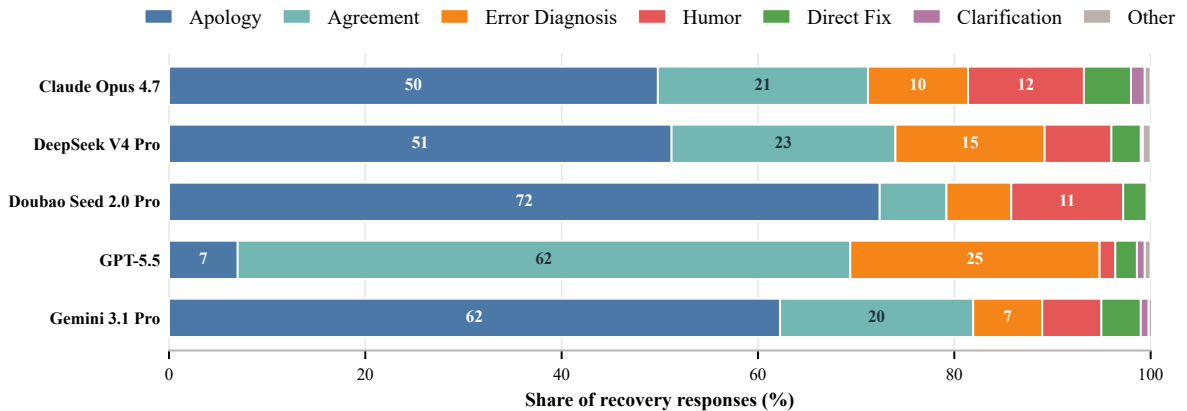


Figure 11: Breakdown of failure recovery strategies in representative models.

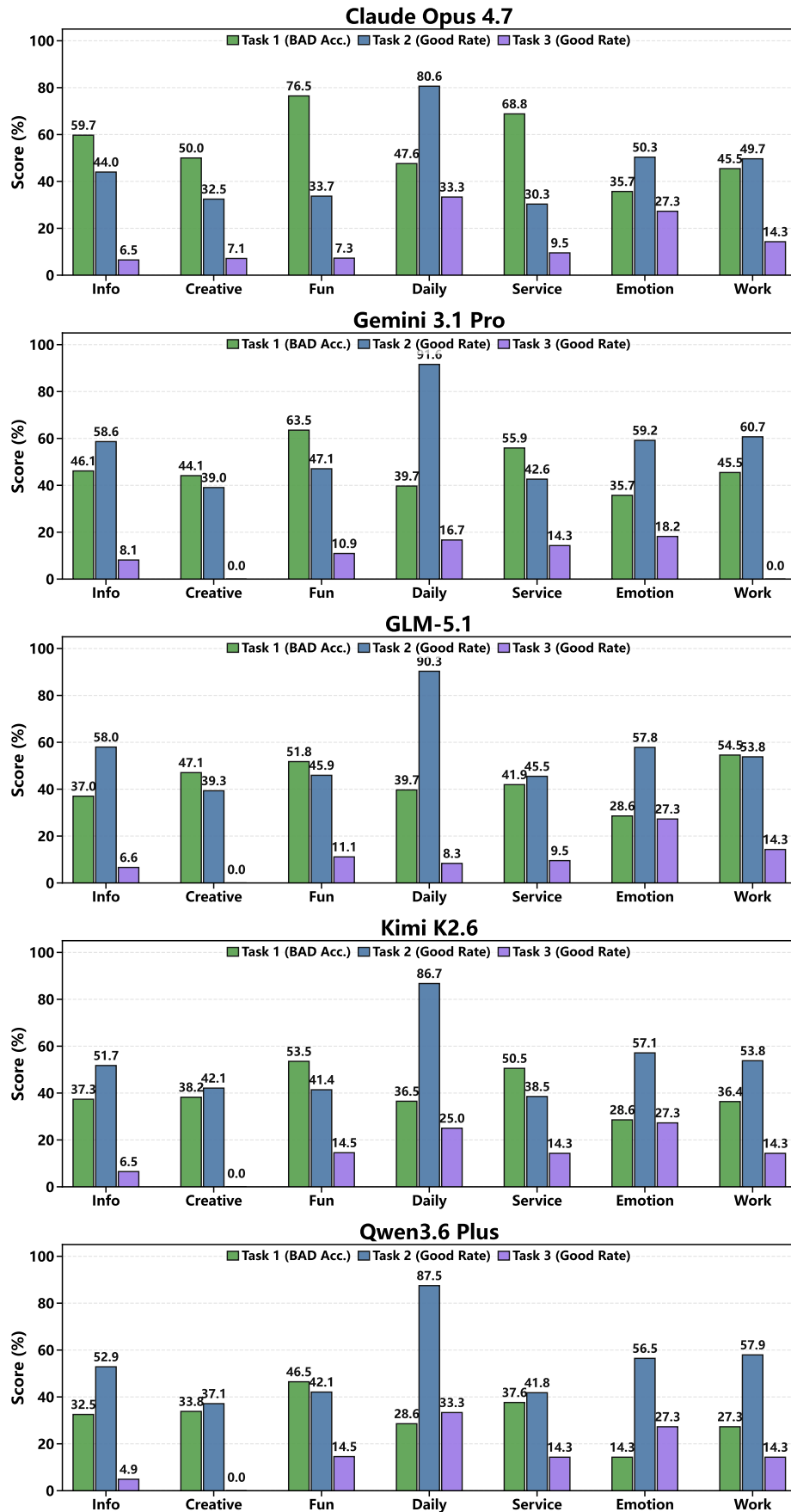


Figure 12: Scenario-level performance across UXBench tasks for five representative models. Each row shows one model, with grouped bars comparing Task 1, Task 2, and Task 3 across scenarios.