CoReflect: Conversational Evaluation via Co-Evolutionary Simulation and Reflective Rubric Refinement

Anonymous ACL submission

Abstract

Evaluating multi-turn conversational systems remains a fundamental challenge. Conventional pipelines rely on manual rubrics and fixed context, a static approach that limits coverage and fails to capture emergent dialogue behaviors. To address this, we introduce CoReflect (Conversational Evaluation via Co-Evolutionary Simulation and Reflective Rubric Refinement), which unifies dialogue simulation and evaluation into an adaptive, iterative process. CoReflect employs a conversation planner that generates structured templates to guide a user simulator through diverse, goal-directed dialogues. Subsequently, a reflective analyzer processes these dialogues to identify systematic behavioral patterns and automatically refine the evaluation rubrics. Crucially, the insights from the conversation analysis are fed back into the planner to update conversation templates for subsequent iterations. This co-evolution loop ensures that the complexity of test cases and the diagnostic precision of rubrics improve in tandem. By minimizing human intervention, CoReflect provides a scalable and self-refining methodology that allows evaluation protocols to adapt alongside the rapidly advancing capabilities of dialogue models.

1 Introduction

Large language models (LLMs) increasingly power multi-turn applications such as personalized tutors and customer service agents (Brown et al., 2020; OpenAI, 2023; Anthropic, 2024b; Comanici et al., 2025). In these dynamic settings, users demand more than linguistic fluency; they expect coherent, adaptive, and contextually grounded dialogues. Evaluating such conversational capabilities, however, remains a challenging open problem.

To date, the most reliable approach to conversational evaluation relies on open-ended human assessment against predefined rubrics, as human raters can dynamically probe nuanced behaviors and value alignment over extended interactions. However, this approach is costly, time-consuming, and prone to subjective bias. While automated alternatives, such as "LLM-as-a-judge" frameworks, offer efficiency and have shown promising alignment with human judgments (Liu et al., 2023b; Zheng et al., 2023; Zhao et al., 2025), they often rely on static, pre-designed conversational contexts that fail to capture authentic model behavior. To address this, researchers have turned to LLM-based simulators to generate more dynamic, fluent, and natural conversational interactions with test chatbots. However, these simulators often struggle to systematically elicit challenging or adversarial behaviors, and may exhibit shallow exploration or mode collapse in self-driven interactions (Park et al., 2023; Shanahan et al., 2023; Chiang et al., 2024; Wei et al., 2025).

More fundamentally, both human and automated evaluation approaches can be constrained by the evaluation architect's subjective focus and limited foresight. Overly narrow rubrics may fail to capture emergent model capabilities or unforeseen behaviors that manifest during subsequent deployment, leading to costly redesign and re-execution of the evaluation process.

To address these limitations, we propose **CoReflect**, a co-evolutionary and reflective evaluation framework that moves beyond static benchmarks and fixed rubrics. By introducing an adaptive, iterative process, CoReflect ensures that both dialogue simulation and evaluation criteria evolve in tandem to capture model behaviors. Instead of relying on static conversational scripts, the framework utilizes a *conversation planner* as an engine that orchestrates the entire interaction under specific user persona and scenario. Specifically, the planner generates structured, goal-directed templates that prescribe specific turn-level instructions for a user simulator on how to interact with test models.

These resulting interactions are then assessed by

an *LLM-as-a-judge evaluator* using a structured three-step protocol—analyzing turns, synthesizing observations, and assigning ratings against the current rubrics. Following the evaluation results, a *reflective analyzer* processes these conversation assessments to cluster behavioral patterns and systematic failures, extracting insights to refine the evaluation rubrics automatically. These insights are also fed back to the planner, which updates its conversation templates to target identified weaknesses with greater precision in subsequent iterations.

Our main contributions are presented as follows:

- **Co-evolutionary conversational evaluation framework.** We propose CoReflect, a novel co-evolutionary evaluation framework in which conversation templates and evaluation rubrics are jointly refined through an iterative feedback loop. This co-evolution mechanism is the core innovation of our work, allowing the user simulation and the diagnostic precision of the rating criteria improve in tandem.

- **Autonomous synthesis of evaluation rubrics.** CoReflect introduces a reflective analyzer as the primary mechanism for translating model behavioral data into structured evaluation logic, effectively alleviating the human burden of manual design. By autonomously synthesizing patterns from simulated interactions, it mitigates subjective human bias and enables the system to dynamically expand its scope beyond static rubrics. This data-driven approach can capture emergent model behaviors that might otherwise remain hidden, ensuring the evaluation framework evolves in lockstep with the model's complexity.

- **Large-scale conversational evaluation.** We conduct extensive automatic and human-in-the-loop experiments across diverse personas, scenarios, interaction categories, and model families, demonstrating that CoReflect enables fine-grained and consistent evaluation of state-of-the-art LLMs in multi-turn conversational settings.

2 Methodology

The evaluation process is initialized by a human architect who manually defines a set of coarse, high-level rubrics to serve as the foundational criteria for assessing the conversational quality of several test models. In each iteration, each test model interacts with a user simulator, which is guided by *a conversation planner*, across various personas and scenarios. This planner generates goal-directed, persona-grounded conversation templates that instructs the user simulator on how to conduct each user turn, thereby precisely challenging the models against the given rubrics.

An evaluator assesses these simulated conversations against the established rubrics, providing quantitative ratings and qualitative rationales. These results are then sampled and processed by *a reflective analyzer* to characterize model behaviors ad limitations, offering insights that drive an iterative refinement loop; in this loop, the rubrics are tuned and the planner accordingly updates its templates to better challenge the models in subsequent rounds. Figure 1 illustrates this interplay between the planner, evaluator, and reflective analyzer. This process proceeds through time steps $t = 1, \ldots T$, where $t = 1$ represents the initial state, and terminates after T iterations. The specifics of the framework and the process are provided below:

2.1 Personas, Scenarios, and Test Models

Each evaluation *instance* in this study comprises two parts: (i) a user, who is manifested through a persona profile, (ii) a scenario, which defines a context, task, and objective of the conversation. We next detail the persona and scenario construction process used to generate an evaluation set.

Personas. To capture user behaviors in the evaluation, we create structured personas that specify how users express themselves and how they expect the system to interact with them. Specifically, each persona comprises two parts: (1) *User traits*: the user's attributes manifested in their communication, such as tone, verbosity, and conversational nuances. We denote user traits by x. (2) *Response preferences*: the desired qualities of the model's responses, including reasoning depth, level of detail, and formality, etc. We denote user preferences by θ. This representation enables controlled variation across communication styles while maintaining realism and interpretability. Complete schema definitions and examples are provided in Appendix A.

Scenarios. Scenarios are created independently from personas to ensure broad coverage of conversational goals. Each scenario specifies a context, primary task, objective of the conversation, and an estimate of number of turns to complete the task. We categorize scenarios into four intent types–*instructional*, *informational*, *operational*, and *interactive*, as detailed in Appendix B. Collectively,

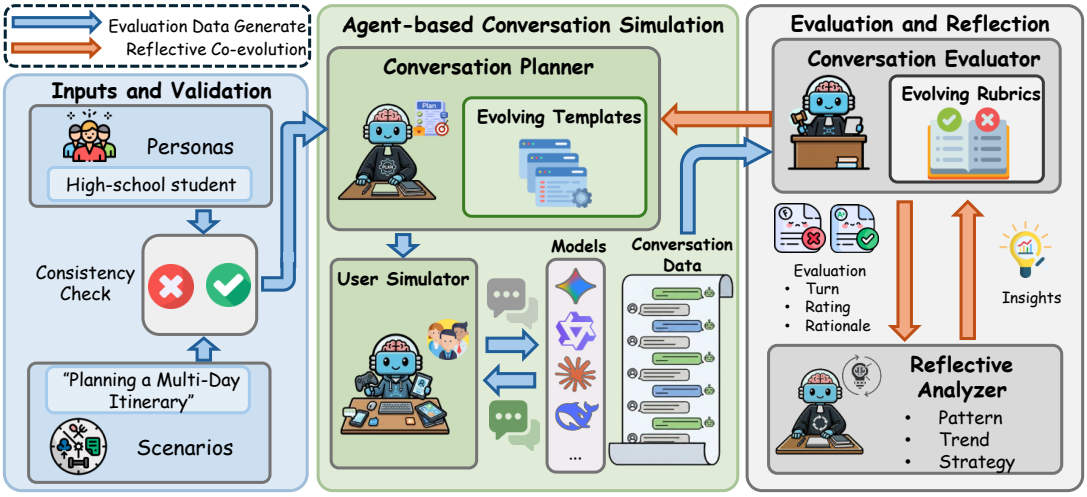


Figure 1: CoReflect overview. Initially, evaluation instances are generated from persona-scenario pairs and validated to ensure contextual coherence. Then the co-evolution loop proceeds in two steps: (i) *Conversation simulation*, where a planner generates structured templates to guide a user simulator through goal-directed dialogues across diverse persona-scenario pairs, and (ii) *Reflective rubric refinement*, where an LLM-as-a-judge evaluates the interactions and a reflective analyzer extracts insights from clustered behavioral patterns to update the rubrics. These insights feed back into the planner to update conversation templates for subsequent iterations.

these categories encompass factual, task-oriented, and social interactions. We denote a scenario by s.

Pairing process and consistency check. Each persona is paired with every scenario. However, not every persona–scenario combination yields a contextually-coherent setup (e.g., a young student tutoring a senior engineer). Each persona–scenario pair undergoes a consistency check by a model-based verifier to validate its contextual coherence. Only those pairs validated as plausible are retained. Let $\mathcal{D}$ denote this dataset of validated persona–scenario pairs and we write the i-th pair as a tuple (x_i, θ_i, s_i). See Appendix C for more details.

Test models. Each test model is evaluated against the evaluation set $\mathcal{D}$. The collection of test models is denoted by $\mathcal{M} \triangleq \{m_j\}_{j=1}^{M}$, where each m_j represents a specific conversational model and $M = |\mathcal{M}|$ is the total number of the models.

Rubrics. We define the set of rubrics at the start of iteration t by $\mathcal{R}^{(t)} \triangleq \{R_k^{(t)}\}_{k=1}^{K}$, where K denotes the total number of rubrics in the evaluation. While the initial set $\mathcal{R}^{(1)}$ is manually drafted, the rubrics are iteratively refined in subsequent steps.

2.2 Planner and Conversation Simulation

At iteration t, the planner determines the required turn count $N_i^{(t)}$ for each instance $(x_i, \theta_i, s_i) \in \mathcal{D}$ and generates a structured conversation template $\mathcal{T}_i^{(t)}$ (schema in Appendix D). These turn-level instructions guide the dialogue to unfold naturally while systematically probing specific model capabilities like reasoning and clarification. To continuously improve template quality, this generation process incorporates insights from the preceding iteration, $\mathcal{I}^{(t-1)}$ (where $\mathcal{I}^{(0)} \triangleq \emptyset$; derivation details are in Section 2.4).

Once the templates for all persona–scenario pairs in $\mathcal{D}$ are created, each model $m_j \in \mathcal{M}$ conducts a dialogue with the user simulator for every pair. The simulator is provided with the instance (x_i, θ_i, s_i) and the corresponding template $\mathcal{A}_i^{(t)}$, which serve as contextual grounding and turn-level instructions. In contrast, model m_j has only user traits x_i as input information. The simulator and model m_j then have a dialogue, based on their respective information sets and the evolving conversation history. This asymmetric information is designed to simulate real-world chatbots leveraging historical personal data (i.e., x_i) to personalize interactions. We denote the resulting conversation generated by model m_j for the i-th instance at iteration t as $C_{i,j}^{(t)}$.

2.3 Conversation Evaluator

An LLM-as-a-judge evaluates each generated conversation $C_{i,j}^{(t)}$ by model m_j against the corresponding instance (x_i, θ_i, s_i) and the rubric set $\mathcal{R}^{(t)}$.

The judge generates a rating and a correspond-

ing rationale for each rubric, following a structured three-step reasoning protocol. First, it performs a turn-level analysis to generate concise observations for every model response grounded in the immediate context. Next, it aggregates these turn-level observations to identify conversation-wide strengths and weaknesses. Finally, the evaluator assigns a numerical score to each rubric alongside a rationale that references the synthesized evidence.

Let $r_{i,j,k}^{(t)}$ and $\rho_{i,j,k}^{(t)}$ denote the rating and rationale for conversation $C_{i,j}^{(t)}$ as evaluated under rubric $R_k^{(t)} \in \mathcal{R}^{(t)}$. The performance of model m_j for a specific rubric $R_k^{(t)}$ is defined as the mean rating across the conversation set $\mathcal{D}$; the model's aggregate performance, $\mu_j^{(t)}$, is then calculated as the average across all K rubrics:

$$\mu_{j,k}^{(t)} \triangleq \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} r_{i,j,k}^{(t)}, \quad k = 1, \ldots, K$$

$$\mu_j^{(t)} \triangleq \frac{1}{K} \sum_{k=1}^{K} \mu_{j,k}^{(t)}.$$

2.4 Reflective Analyzer for Rubric Refinement and Planner Feedback

A reflective analyzer synthesizes the collected ratings $\{r_{i,j,k}^{(t)}\}$ and rationales $\{\rho_{i,j,k}^{(t)}\}$ to extract insights $\mathcal{I}^{(t)}$ regarding model behavior. These insights are then leveraged to sharpen the rubrics' diagnostic precision and to refine the planner's templates. The reflective analyzer proceeds as follows:

(1) **Instance sampling:** Based on the ratings $\{r_{i,j,k}^{(t)}\}$, we partition the conversations into high- and low-rated tiers. To ensure their representation in the analysis, we sample equally from each tier to construct an analysis pool, retaining only the associated rationales $\rho_{i,j,k}^{(t)}$ as structured evidence for behavioral analysis.

(2) **Pattern discovery:** The sampled rationales are embedded and clustered into a set of *behavioral families*, $\mathcal{F}^{(t)} \triangleq \{f_\ell^{(t)}\}_{\ell=1}^{L^{(t)}}$, where $L^{(t)}$ represents the number of distinct, recurring patterns identified at iteration t. Each family f_ℓ characterizes a specific conversational pattern, including both undesirable behaviors (e.g., loss of task focus, stylistic drift) and effective strategies (e.g., proactive clarification, structured task decomposition).

(3) **Insight synthesis and rubric refinement.** For every family $f_\ell^{(t)}$, the reflective analyzer produces an interpretable insight $\iota_\ell^{(t)}$ that characterizes the model behavior and specifies the criteria for reward or penalty. These individual insights are aggregated into a comprehensive set $\mathcal{I}^{(t)} \triangleq \{\iota_\ell^{(t)}\}_{\ell=1}^{L^{(t)}}$, which is utilized to update the performance definitions, rating anchors, and evidence cues of the relevant rubrics in $\mathcal{R}^{(t)}$. This refinement enhances the framework's diagnostic precision for the subsequent iteration, formalized as:

$$\mathcal{R}^{(t+1)} \leftarrow \text{UPDATE}\Big(\mathcal{R}^{(t)}, \mathcal{I}^{(t)}\Big)$$

The derived insights $\mathcal{I}^{(t)}$ feed into the conversation planner, refining the templates to target identified behaviors more precisely in the subsequent iteration.

2.5 Measures on Rubrics Refinement

To assess how iterative rubric refinement affects evaluation quality, we track two complementary metrics across iterations: rubric *discriminability* and *stability*. Discriminability measures the extent to which the evaluator separates models with different capability levels, operationalized as the inter-model variability of mean ratings $\sigma_{\text{inter}}(t)$. Stability quantifies the evaluation consistency through intra-model score variance $\sigma_{\text{intra}}^2(t)$ and rank consistency measured by Spearman's $\rho(t)$. Their formal definitions are provided in Appendix E.

3 Experiments

3.1 Experiment Setup

Datasets. We created 30 personas, each embodying distinct communication styles and response preferences, and 10 scenarios per category (cf. Table 5), yielding 953 validated persona-scenario pairs after automated consistency checks. Summary statistics are provided in Table 1.

Test models. We evaluated seven frontier LLMs from different providers: Gemini-2.5-pro, Gemini-

Category	#persona	# validated pairs	Avg. required turns per template
Informational	30	215	6.60
Instructional	30	244	6.97
Interactive	30	273	7.23
Operational	30	221	6.48
Total	**30**	**953**	**6.85**

Table 1: Summary statistics of the evaluation dataset.

2.5-flash (Comanici et al., 2025); Claude Sonnet-4.5, Claude Haiku-4.5 and Claude Sonnet-4 (Anthropic, 2024b,a); Qwen3-Next (Yang et al., 2025), DeepSeek-R1-671B (Guo et al., 2025).

Rubrics design. We evaluated model performance with $K = 6$ rubrics under two high-level dimensions: *Task Completeness (TC)*, which measured whether the model understood the task intent and makes coherent progress toward the scenario objective, and *User-Centric Personalization (UCP)*, which measured whether the model adapted to the user's preferences, identity, and interaction style throughout the dialogue. Each rubric was rated on a 1–5 scale, where 5 denoted the ideal outcome. While the rubric names were fixed, their descriptions and rating guidelines were iteratively refined through the co-evolution loop over $T = 3$ iterations. The complete rubric definitions and their rating criteria are provided in Appendix F; implementation details are provided in Appendix G.

Human validation. We conductd two human validation studies to evaluate (i) alignment between the LLM-based evaluator and human judgment, and (ii) the quality of the user simulator.

For evaluator validation, we sampled 40 persona-scenario pairs (10 per interaction category). Three annotators rated the evaluate conversations across all three co-evolution rounds using the corresponding rubrics and five-point criteria. We then compared the LLM-as-a-judge scores directly against these human ratings. We reported MAE for absolute score discrepancy with respect to averaged human ratings, and quadratic weighted kappa (QWK) for ordinal agreement on the 1–5 rubric scale. For QWK, agreement was computed separately with each annotator and then averaged, avoiding additional discretization of aggregated human scores.

In addition, we conducted a small-scale human assessment of the user simulator to evaluate its quality and naturalness of generated interactions. Details of this study are provided in Appendix H.

3.2 Experiment Results and Analysis

3.2.1 Model Performance and Stratification

Table 2 details the final-round ($T = 3$) ratings for multi-turn task execution and personalization, showing a clear stratification of model capabilities.

Gemini 2.5 Pro leads the frontier with the highest model rating of 4.81, dominating all evaluation rubrics. This performance reflects that the model's superior reasoning and planning capabilities translate into highly effective dialogue management over complex, multi-turn interactions.

By contrast, the other models show mixed performance, excelling in specific areas while lagging in others. This variance underscores that high-quality multi-turn interaction requires more than isolated strengths; it demands a balanced integration of understanding, task progression, and user adaptation.

Task Completeness: understanding versus execution. *Gemini 2.5 Pro* achieves the highest overall Task Completeness with an average score of 4.81, outperforming *DeepSeek-R1* (the best non-Gemini baseline, 4.64) by 3.7% This corresponds to a relative improvement of approximately 3.7% over the best non-Gemini baseline (*DeepSeek-R1*, 4.64), and smaller models like *Claude Haiku 4.5* by over 22%. Crucially, *Gemini 2.5 Pro* also leads in Functional Task Progression (FTP), demonstrating superior task decomposition and state tracking across long-horizon interactions.

By contrast, runner-up models exhibit clear trade-offs between conceptual understanding and execution. For instance, *the Claude Sonnet series* and *DeepSeek-R1* achieve strong Domain Conceptual Alignment (DCA), yet their FTP scores lag behind *Gemini 2.5 Pro* by 25% or more in relative terms (e.g., *Claude Sonnet 4.5*). Conversely, *Qwen3-Next* prioritizes execution, with an FTP score within 4% of *Gemini 2.5 Pro*, but at the cost of substantially weaker DCA (approximately 17% lower), suggesting brittle execution under conceptual ambiguity. This execution gap is further magnified in reduced-capacity variants: *Gemini 2.5 Flash* and *Claude Haiku 4.5* trail *Gemini 2.5 Pro* in FTP by roughly 12% and 33% respectively, highlighting the difficulty of sustaining long-horizon task progression without sufficient model capacity.

User-Centric Personalization: style consistency versus adaptability. *Gemini 2.5 Pro* leads in User-Centric Personalization with a 4.80 average score, maintaining a consistent 5% relative advantage over its closest competitors. Its advantage is most pronounced in Anticipatory Flow Management (AFM), where it outperforms *Claude Sonnet 4.5* by over 15% and *Claude Haiku 4.5* by roughly 6%, demonstrating its capability to anticipate evolving user intent rather than merely reacting to explicit cues. Meanwhile, *Claude Sonnet 4*, *DeepSeek-R1*, and *Qwen3-Next*, cluster closely in

Test models	Task Completeness				User-Centric Personalization				Model rating
	ODI	DCA	FTP	**avg.**	AFM	OSF	SSA	**avg.**	
Claude Sonnet 4	4.65	4.76	4.25	4.56	4.75	4.74	4.73	4.74	4.65
Claude Sonnet 4.5	4.37	4.69	3.53	4.20	4.13	4.68	4.57	4.46	4.33
Claude Haiku 4.5	4.02	4.64	3.16	3.94	4.47	4.57	4.61	4.55	4.25
Qwen3-Next	4.73	4.03	4.50	4.42	4.62	4.72	4.70	4.68	4.55
DeepSeek-R1	4.68	4.65	4.59	4.64	4.60	4.54	4.54	4.56	4.60
Gemini 2.5 Pro	**4.86**	**4.86**	**4.70**	**4.81**	**4.76**	**4.86**	**4.79**	**4.80**	**4.81**
Gemini 2.5 Flash	4.79	4.72	4.12	4.60	4.70	4.72	4.76	4.75	4.68

Table 2: Model performance across all rubrics at iteration $t = 3$. (ODI: Output Delivery Integrity, DCA: Domain Conceptual Alignment, FTP: Functional Task Progression; AFM: Anticipatory Flow Management, OSF: Output Structure Fit, SSA: Sustained Style Adherence). See Appendix J for results at $t = 1, 2$.

Metric	$t = 1$	$t = 2$	$t = 3$
Rubric Discriminability			
Inter-model std. dev. ($\sigma_{\text{inter}}(t)$)	0.062	0.128	**0.194**
Rubric Stability			
Intra-model variance ($\sigma^2_{\text{intra}}(t)$)	0.145	0.141	**0.138**
Rank consistency (Spearman $\rho(t)$)	0.68	0.79	**0.92**

Table 3: Discriminability and stability across reflection iterations. Iterative refinement increases inter-model separation while preserving evaluation consistency.

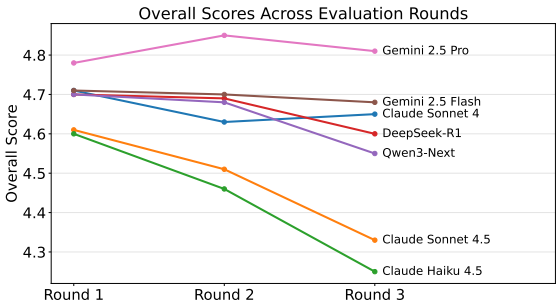


Figure 2: Model ratings across iterations. Iterative reflection improves model differentiation, yielding more distinct stratification and relative ordering.

overall personalization performance, with performance variance staying within 5%.

In contrast, Claude Sonnet 4.5 and Claude Haiku 4.5 match the baseline on stylistic consistency (OSF/SSA scores within 5%) but falter on adaptability, with Sonnet's AFM performance lagging by more than 13%. This divergence highlights a distinct architectural trade-off: while stylistic alignment is highly robust across frontier models, the capacity to steer conversations as user intent shifts remains strictly dependent on model capacity.

3.2.2 Effect of Co-Evolution Iterations

Figure 2 shows that model ratings are highly clustered in the initial round, limiting reliable model ordering. After iterative reflective rubric refinement and co-evolutionary user simulation, later iterations reveal more fine-grained stratification of model performance, suggesting that CoReflect progressively uncovers subtle quality differences that are initially obscured. This trend is quantitatively supported by Table 3: inter-model rating dispersion increases monotonically across iterations, while intra-model variance remains stable and rank consistency improves. These results indicate that CoReflect strengthens evaluative signal without introducing additional stochastic noise.

Table 4 presents our human validation study, showing that evaluator-human alignment improves across co-evolution rounds. The largest gain occurs between the first and second rounds, followed by more incremental improvements in the third. This trend suggests that early reflection substantially corrects coarse evaluator misalignment, while subsequent iterations yield finer refinements.

However, these gains are not uniform across rubrics. Structurally sensitive dimensions (OSF) and task-progression dimensions (FTP) exhibit the largest reductions in MAE. Conversely, DCA begins with stronger initial agreement, leaving less room for improvement. Overall, MAE decreases from 0.70 to 0.51 and QWK increases from 0.44 to 0.59, demonstrating that CoReflect substantially enhances both score calibration and ordinal agreement with human judgments.

3.2.3 Effect of Conversation Length

Figure 3 illustrates the impact of conversation length on model performance. Across all models, both Task Completeness and User-Centric Personalization scores decline as conversation length increases, indicating that extended interactions pose a systematic challenge for current models.

Round	Metric	Task Completeness				User-Centric Personalization				Overall
		ODI	DCA	FTP	**avg.**	AFM	OSF	SSA	**avg.**	
1st-round	MAE ↓	0.68	0.54	0.76	0.66	0.71	0.83	0.69	0.74	0.70
	QWK ↑	0.46	0.58	0.39	0.48	0.42	0.34	0.45	0.40	0.44
2nd-round	MAE ↓	0.53	0.49	0.58	0.53	0.56	0.61	0.55	0.57	0.55
	QWK ↑	0.58	0.62	0.51	0.57	0.54	0.52	0.56	0.54	0.56
3rd-round	MAE ↓	**0.48**	**0.47**	**0.52**	**0.49**	**0.51**	**0.55**	**0.52**	**0.53**	**0.51**
	QWK ↑	**0.62**	**0.63**	**0.56**	**0.60**	**0.58**	**0.57**	**0.58**	**0.58**	**0.59**

Table 4: Human validation of LLM-as-a-judge evaluator across iterations. MAE measures absolute discrepancy against averaged human ratings while QWK measures ordinal agreement with individual human annotators. Lower MAE and higher QWK indicate stronger alignment with human judgments.

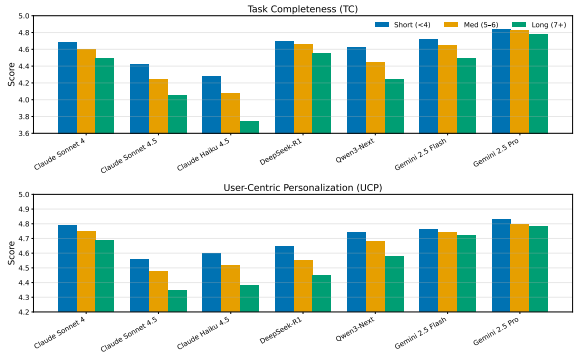


Figure 3: Model ratings of Task Completeness and Personalization, stratified by conversation length.

Task completeness degrades more significantly than personalization in long-horizon dialogues. While the models maintain stable personalization, their ability to track task states falters over time. This trend widens the performance gap between large-scale and lightweight models, which is indistinguishable in shorter interactions.

These results suggest conversation length as a critical factor for long-horizon reasoning and execution, highlighting the necessity of length-aware analysis in multi-turn conversational benchmarks to accurately differentiate model capabilities.

3.2.4 Case Study of Rubric Refinement

Figure 4 illustrates an example of reflective rubric refinement: A test model is requested to present intermediate variables and results in a standardized table format. While the model initially follows the required structure, it later reverts to inline narrative text in a subsequent turn. Similar structural regressions were repeatedly observed across diverse multi-turn scenarios.

The reflective analyzer identifies these patterns and refines the initial rubric to capture cross-turn structural consistency as an important dimension. Specifically, in this example, the *Output Structure Fit* rubric is updated from emphasizing general formatting quality to explicitly evaluating alignment with user preferences and previously established output schemas. The five-point scoring criteria are also refined to better distinguish varying levels of structural consistency (See Appendix J for details).

4 Related Work

4.1 Structured User Simulation

Structured simulation provides a controllable framework for systematically probing model behavior, ranging from classical planning-based systems (Reiter and Dale, 2000) to modern neural plan-and-write pipelines (Yao et al., 2019; Xu et al., 2020). In conversational settings, these principles have been extended to LLM-based user simulators, which enable interactive evaluation beyond fixed contexts. For example, Wang et al. (2023) employ a ChatGPT-based simulator for multi-turn conversational recommendation evaluation, while Liu et al. (2023a) show that using multiple simulators improves robustness compared to relying on a single simulator. More broadly, persona-driven and template-based simulations enable targeted stress-testing of model behaviors such as personalization, consistency, and long-horizon task execution.

4.2 LLM-as-a-Judge and Scalable Evaluation

LLM-as-a-judge has emerged as a widely adopted paradigm to scale evaluation beyond human annotation. G-Eval (Liu et al., 2023b) demonstrates that LLM-based evaluators can achieve strong correlation with human judgments, while MT-Bench (Zheng et al., 2023) and CHATBOT ARENA (Chiang et al., 2024) enable large-scale

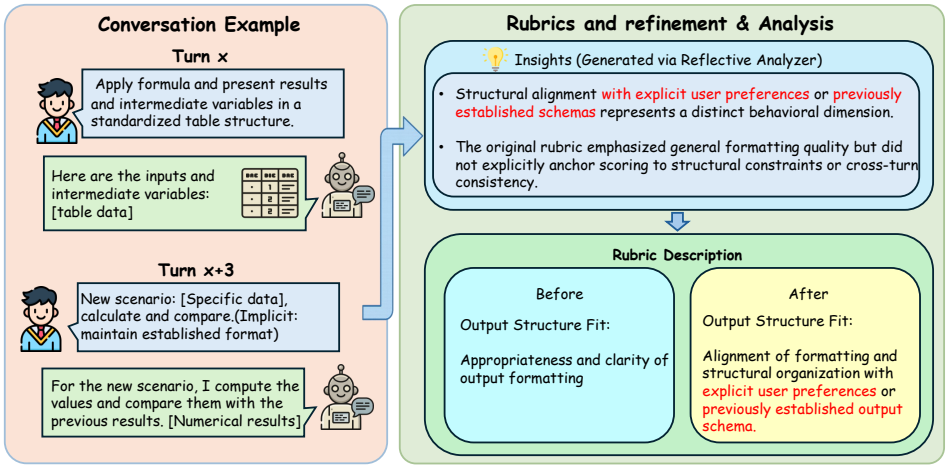


Figure 4: Example of reflective rubric refinement for cross-turn structural regressions. A test model initially follows a user-requested table schema but later reverts to inline narrative text. The reflective analyzer identifies the failures, updating *Output Structure Fit* to evaluate alignment with explicit user preferences or established output structures.

multi-turn comparison through pairwise or rubric-based scoring. Complementary benchmarks such as LongBench (Bai et al., 2024), L-Eval (An et al., 2024), and RULER (Hsieh et al., 2024) focus on long-context understanding and sustained reasoning. Despite their effectiveness, most LLM-based evaluation frameworks use static rubrics or context, making them less sensitive to new failure modes in personalized and long-horizon interactions.

4.3 Self-Refining Evaluation

To address the limitations of static LLM-as-a-judge frameworks, recent work explores *self-refining* evaluation paradigms using reflection and iterative feedback loops. One approach employs reflective evaluation, where LLMs generate structured rationales or critiques before scoring to surface latent failure modes (Liu et al., 2023b; Zheng et al., 2023; Shinn et al., 2023). Building on such signals, other work introduces iterative refinement loops that update evaluation prompts, rubrics, or protocols based on low-performing or ambiguous cases. Hu et al. (2023) use LLM-simulated user feedback as an annotation-free signal to iteratively optimize dialogue responses. These studies highlight a shift toward closed-loop frameworks that continuously refine evaluation criteria as model behaviors evolve.

4.4 Evaluation of Personalized Dialogue

Personalization in conversational AI has centered on persona grounding, preference modeling, and adaptive generation. Foundational benchmarks like PERSONA-CHAT (Zhang et al., 2018), BLENDED-SKILLTALK (Smith et al., 2020), and EMPATHETICDIALOGUES (Rashkin et al., 2019) established controlled environments for assessing consistency, multi-skill agility, and emotional grounding. Recent efforts, such as DialogBench (Ou et al., 2024) and PersonaConvBench (Li et al., 2025), have extended these evaluations to multi-turn interactions. However, while these resources effectively identify personalization failures, they rely on static scripts and evaluation criteria that struggle to adapt as conversational behaviors and user expectations evolve.

5 Conclusion

We introduced CoReflect, an autonomous framework unifying persona-driven simulation with iterative rubric refinement. By dynamically evolving evaluation criteria based on observed behaviors, CoReflect captured nuances that static benchmarks tend to miss. Experiments confirmed this self-refining approach significantly improved discriminatory resolution, resulting in clear, high-fidelity model stratification.

CoReflect provides a scalable solution for complex, hard-to-define domains such as personalization by automatically updating obsolete criteria and mitigating the subjective bias inherent in human-designed evaluation architectures. Future work will aim to strengthen CoReflect by grounding simulation in real user interactions and improving evaluation robustness through richer, bias-aware judging and rubric design.

Limitations

While CoReflect establishes an adaptive and systematic framework for conversational evaluation, it has two primary limitations. First, the framework relies on simulated user interactions driven by predefined, synthetic personas and scenarios. Although this design ensures control and scalability, these simulations may not fully capture the spontaneity, emotional nuance, and long-term dynamics of authentic human interaction. Second, CoReflect employs an LLM-as-a-judge for both rating and rubrics refinement. As a result, the evaluation process risks inheriting the intrinsic biases or blind spots of the underlying judge model, which could skew the trajectory of rubric evolution. Critically, while this study demonstrates the framework's internal consistency, we have not yet conducted a human assessment to externally validate the generated rubrics or the judge's scoring alignment. Future work must address this by comparing CoReflect's automated outputs with human expert annotations and by employing diverse models as judges to investigate potential self-preference biases.

Ethical Considerations

To validate the realism of our user simulator, we conducted a study with three volunteer annotators under an ethically approved protocol. Participants were fully informed of the study's purpose, and the evaluation was performed on randomly sampled simulated interactions that contain no sensitive information and no real-world human data. Human ratings were collected using a transparent 5-point Likert scale to ensure objective assessment, and all data were used strictly in accordance with their intended purposes. As a general evaluation framework, CoReflect does not explicitly control or filter for potentially offensive or unsafe content generated by the underlying base LLMs. Our objective is to evaluate conversational behaviors as exhibited by the models themselves, rather than to impose additional safety constraints that could confound behavioral analysis. We view content safety and moderation as complementary concerns that can be incorporated through external filters or safety-oriented rubrics in future extensions of the framework.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411.

Anthropic. 2024a. The claude 3 model family: Haiku, sonnet, & opus. Technical report.

Anthropic. 2024b. Claude: An ai system by anthropic. https://www.anthropic.com/claude.

Xiaozhi Bai, Ning Ding, Yulin Chen, Zhengxiao Li, Zhiyuan Liu, and Maosong Sun. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Daniel Li, ..., and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Cheng-Ping Hsieh, Shih-Yang Sun, Samuel Kriman, Suyog Acharya, Debajyoti Rekesh, Fei Jia, Yiming Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*. ArXiv:2404.06654.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3953–3957.

Li Li, Peilin Cai, Ryan A Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, and 1 others. 2025. A personalized conversational benchmark: Towards simulating personalized conversations. *arXiv preprint arXiv:2505.14106*.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023a. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–21, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. 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.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5370–5381.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623:493–498.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, pages 1–15. Curran Associates, Inc.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10052–10065, Singapore. Association for Computational Linguistics.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, and 1 others. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro. 2020. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.

Shangqing Zhao, Ming Hong, Yutao Liu, Devamanyu Hazarika, and Kevin Lin. 2025. Prefeval: Do llms recognize your preferences? evaluating personalized preference following in llms. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Shiyi Zhuang, Zi Lin Wu, Yong Zhuang, ..., and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Appendix

A User Persona Definition

To simulate diverse user profiles in personalized multi-turn conversations, we define a structured persona schema capturing both the user's expressive behavior and their expectations for AI responses. Each persona is composed of the following elements:

ID and Role A unique identifier and a brief description of the user's background or profession (e.g., University Student, Senior Developer). This provides contextual grounding for the user's domain knowledge and communication style.

Language The primary language used by the persona, enabling multilingual dialogue simulation and evaluation.

Personality This component models how the user naturally communicates. It includes:

- **Traits:** Core personality characteristics that influence conversational behavior (e.g., Impatient, Inquisitive, Skeptical).
- **Tone:** The persona's typical emotional tone in conversation (e.g., Formal, Sarcastic, Friendly).
- **Verbosity:** Typical message length and level of detail (e.g., Terse, Verbose).
- **Quirks:** Specific linguistic habits that define the user's style (e.g., Skips pleasantries, Uses all lowercase, Frequently references personal projects).

Preferred Response Style This defines how the user expects the AI assistant to respond and serves as the personalization target. It includes preferences for:

- Tone (e.g., Empathetic, Direct),
- Verbosity (e.g., Concise, Detailed),
- Reasoning Depth (e.g., Brief rationale, Step-by-step explanation),
- Engagement (e.g., Neutral, Asks follow-ups),
- Clarity (e.g., Simple and clear, Technically precise).

B Scenario Category Definition

To comprehensively evaluate the capabilities of large language models in multi-turn dialogue, we define four broad categories of interaction scenarios and a structured schema for their definition.

Scenario Categories

- **Instructional dialogue:** Encompasses interactions where the user's objective is to acquire new knowledge or develop specific skills. The model serves in an instructional capacity, providing pedagogical guidance and step-by-step reasoning.
- **Informational dialogue:** Characterized by the user's intent to obtain accurate, context-relevant information. The model functions as a factual resource, delivering concise and verifiable responses.
- **Operational dialogue:** Includes interactions where the user seeks assistance in completing a concrete task or producing a functional output. The model is employed as a productivity facilitator to generate content or perform structured actions.
- **Interactive dialogue:** Captures scenarios where the user's primary aim is social, emotional, or imaginative engagement. The model adopts a human-like persona, prioritizing coherence, tone adaptation, and creative collaboration.

Scenario Definition Schema

- **ID and title:** A unique scenario identifier and a concise title summarizing the context (e.g., Planning a Weekly Study Schedule, Summarizing Technical Emails).
- **Category:** A high-level classification of the task domain, selected from predefined types such as Work & Productivity, Learning & Study, Creative & Brainstorming, Personal & Planning, and Information & Factual.
- **Situation description:** A neutral, third-person description of the background or external context.
- **Core task:** A focused objective to be accomplished (e.g., Organize a calendar with overlapping deadlines).

Category	User Intent	Assistant Role	Typical Outputs	Example
Instructional	Acquire knowledge or develop skills	Tutor / Instructor	Explanations, practice questions, scaffolds	*"Teach me how neural networks work."*
Informational	Retrieve facts or clarify references	Reference / Expert	Factual answers, summaries, citations	*"What is the GDP of Brazil in 2023?"*
Operational	Complete a concrete task or create content	Productivity Tool	Code, documents, plans, translations	*"Write a weekly report summary for my team."*
Interactive	Engage socially, emotionally, or creatively	Companion / Partner	Stories, dialogues, empathetic replies	*"I feel nervous about my exam—can we talk about it?"*

Table 5: Scenario categories in the evaluation dataset $\mathcal{D}$, defined by user intents, assistant roles, outputs, and examples.

- **Expected interaction:**
  - Turn Complexity: Short (<4 turns), Medium (5–6 turns), or Long (7+ turns).
  - Dialogue Flow Type: Goal-Oriented, Open-Ended, Step-by-Step Guidance, or Collaborative Generation.
- **Success criteria:** A neutral description of what constitutes a successful outcome (e.g., a clear and logically ordered prioritization of incoming requests).

C Automated Consistency Check

A critical validation step is performed before any persona is paired with a scenario. To prevent the creation of illogical or nonsensical conversations, we employ an automated consistency check using a large language model (LLM) configured as a "judge."

Protocol Steps

1. **Framing the prompt:** The function sends a targeted prompt to an LLM, instructing it to act as a "logical reasoning expert" for the specific task of evaluating a pairing's plausibility.
2. **Providing context:** The prompt is populated with the role from the persona definition and the core_task and title from the scenario definition.
3. **Binary decision:** The LLM is strictly constrained to answer with only a single word, "Yes" or "No."
4. **Filtering:** If the LLM returns "Yes," the pair is approved. If "No," the combination is logged and discarded.

D Conversation Simulation and Template

To simulate realistic multi-turn conversations between a user and an AI assistant, we define a structured generation protocol. Each conversation is generated from a persona-scenario pair and is designed to assess both personalization alignment and task adherence.

Protocol Components

- **Persona context:** Full instantiation of tone, verbosity, quirks, and preferences.
- **Scenario context:** Neutral task description from scenario schema.
- **Generation instructions:**
  - `turnType` (Early Turn, Intermediate Turn, Challenging Turn),
  - `turnIntent` (logic-level description of user's aim),
  - `instructionForEval` (rationale for why this turn tests AI behavior).
- **Conversation structure:**
  - Early Turn: Persona initiates task using typical style.
  - Normal Turns: Task progression through elaboration or clarification.

Example of template prompt with persona and scenario context:	
Context block	**ConversationID:** P001_S01 **Persona (P001):** Senior developer; skeptical, impatient, analytical; direct and terse communication; uses technical jargon and frequently challenges the AI's accuracy. **Scenario (S01):** Instructional dialogue explaining photosynthesis; user seeks a clear, step-by-step explanation of inputs, outputs, and the two main stages (light-dependent and light-independent reactions).
Template	**Turn 1 (Early Turn)** **TurnIntent:** Initiate Core Task **Instruction:** Evaluate the AI's ability to provide a structured, high-level overview and a specific chemical formula as requested. The response should be direct and technical. **Turn 2 (Normal Turn)** **TurnIntent:** Request Specific Detail **Instruction:** Assess the AI's ability to explain a sub-process involving ATP and NADPH with technical accuracy while remaining concise. **Turn 3 (Normal Turn)** **TurnIntent:** Challenge for Deeper Detail **Instruction:** Test the AI's understanding of RuBisCO and carbon fixation, ensuring precision in the biochemical description. **Turn 4 (Preference Recall Turn)** **TurnIntent:** Request Summary & Recall Preference **Instruction:** Check whether the AI recalls the explicit preference for the "balanced chemical equation" stated in Turn 1 and can summarize the steps coherently. **Turn 5 (Normal Turn)** **TurnIntent:** Validate Accuracy with Edge Case **Instruction:** Evaluate whether the AI correctly handles a nuanced biological question involving plant physiology without oversimplification.
Usage	This template illustrates how persona traits and scenario context guide turn-level evaluation. Each turn specifies an intent and an instruction for evaluation, enabling fine-grained assessment of AI behavior in personalized multi-turn dialogue.

Table 6: Example template prompt used in personalized multi-turn conversation evaluation, incorporating essential persona and scenario context.

E Measures for Rubrics Refinement

We formalize the concepts rubric discriminability and stability to quantify the effect of rubric refinement on evaluation quality.

We first define the conversation-level rating $\tilde{r}_{i,j}^{(t)}$ for each $C_{i,j}$ as the average rating across all K rubrics, i.e., $\tilde{r}_{i,j}^{(t)} \triangleq \frac{1}{K}\sum_{k=1}^{K} r_{i,j,k}^{(t)}$. We then compute the variance of these ratings across the conversations $\mathcal{D}$:

$$\sigma_j^2(t) \triangleq \frac{1}{|\mathcal{D}|-1}\sum_{i=1}^{|\mathcal{D}|}\big(\tilde{r}_{i,j}^{(t)} - \mu_j^{(t)}\big)^2,$$

This quantifies the stability of model m_j across diverse scenarios, where a lower variance indicates more predictable performance.

Rubric discriminability. At iteration t, rubric discriminability is defined as the standard deviation of models' mean ratings:

$$\sigma_{\text{inter}}(t) \triangleq \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(\mu_j^{(t)} - \bar{\mu}^{(t)}\right)^2},$$

where $\mu_j^{(t)}$ denotes the mean rating of model j, and

$$\bar{\mu}^{(t)} \triangleq \frac{1}{M}\sum_{j=1}^{M}\mu_j^{(t)}$$

is the average rating across all M models. Higher values of $\sigma_{\text{inter}}(t)$ indicate stronger separation between model capabilities.

Rubric stability. We assess rubric stability using two complementary metrics.

(i) Intra-model variance. We define intra-model variance as the average within-model variance of ratings across conversations:

$$\sigma_{\text{intra}}^2(t) \triangleq \frac{1}{M}\sum_{j=1}^{M}\sigma_j^2(t).$$

(ii) Rank consistency. We compute the Spearman rank correlation coefficient ρ to measure the consistency of model rankings across different evaluation subsets. Higher ρ values indicate that the rubric preserves the relative performance ordering of models despite variations in conversation content.

F Rubrics Design

We evaluate model performance along two high-level dimensions: (i) *Task Completeness (TC)*, which captures whether a model correctly understands task intent and makes coherent progress toward fulfilling the scenario objective; and (ii) *User-Centric Personalization (UCP)*, which reflects the model's ability to adapt its responses to a user's identity, preferences, and interaction style over the course of a dialogue.

Each dimension is assessed using three complementary rubrics, yielding a total of $K = 6$ rubrics:

- **Task Completeness**: Domain Conceptual Alignment (DCA), Output Delivery Integrity (ODI), and Functional Task Progression (FTP) evaluate whether responses are conceptually correct, actionable, and incrementally advance the task.
- **User-Centric Personalization**: Anticipatory Flow Management (AFM), Output Structure Fit (OSF), and Sustained Style Adherence (SSA) assess whether the model anticipates user needs, adapts response organization, and maintains consistent stylistic alignment.

Each rubric is rated on a 1–5 scale, where higher scores indicate better performance. The rubric names remain fixed across iterations to ensure comparability, while their descriptions and rating guidelines are refined through the co-evolution loop. In our experiments, we conduct $T = 3$ iterations.

Below are the refined rubrics by the reflection analyzer at $T = 3$:

Task Execution Refined Sub-rubrics

- **Domain Conceptual Alignment:** Correct identification and sustained adherence to the task's conceptual domain (e.g., fictional vs. real, technical vs. general).
- **Output Delivery Integrity:** Reliable, complete delivery of required artifacts without truncation/corruption (how it delivers, not what).
- **Functional Task Progression:** Multi-step advancement from understanding to precise, actionable outputs.

Personalization Refined Sub-rubrics

- **Sustained Style Adherence:** Consistency in maintaining adapted style and honoring negative constraints across turns.
- **Output Structure Fit:** Alignment of formatting/structure with user preferences or requested schema.
- **Anticipatory Flow Management:** Proactive pacing and next-step guidance that matches user needs and context.

We provide the 5-point refined scoring rules for all the rubrics as follows:

Task Completeness Scoring Rules

- 5 (Exceptional): Perfect task completion, honoring all constraints, with no filler or irrelevant content.
- 4 (Strong): Task met, but with one issue—e.g., slight verbosity, missed minor constraint.
- 3 (Moderate): Goal reached but required user correction or included multiple avoidable errors.
- 2 (Weak): Task only partially met—critical information missing or misunderstood.
- 1 (Poor): Task clearly failed or severely misinterpreted.

Domain Conceptual Alignment

- 5: Perfectly identifies and operates within domain.
- 4: Good alignment, one minor self-corrected drift.
- 3: Some alignment but with inconsistencies.
- 2: Frequent misalignments or wrong domain.
- 1: Complete failure to align with domain.

Output Delivery Integrity

- 5: Flawless and complete delivery with no truncation.
- 4: Full output with one minor glitch.
- 3: Mostly complete but some truncation or errors.
- 2: Significant truncation or multiple errors.
- 1: Severely truncated or failed delivery.

Functional Task Progression

- 5: Autonomously drives task, flawless synthesis into actionable output.
- 4: Good progression with one minor error.

- 3: Progresses task but needs nudging and has errors.
- 2: Requires constant prompting; output has major errors.
- 1: Fails to progress task.

User-Centric Personalization Scoring Rules

- 5 (Exceptional): Flawless personalization—highly adaptive and anticipatory with no missed cues or lapses.
- 4 (Strong): Good personalization but with one noticeable miss (e.g., slight verbosity, delayed adaptation).
- 3 (Moderate): Some effort at personalization, but with inconsistencies or over-reliance on user correction.
- 2 (Weak): Minimal personalization; the model sounds generic, forgets earlier info, or needs multiple corrections.
- 1 (Poor): No sign of personalization—repetitive, forgetful, or misaligned responses.

Sustained Style Adherence

- 5: Flawlessly maintains the adapted style and all negative constraints throughout.
- 4: Maintains style well, with only one minor slip.
- 3: Shows effort but style is inconsistent or requires user correction.
- 2: Frequently deviates or repeatedly violates constraints.
- 1: No sustained adherence; style is erratic.

Output Structure Fit

- 5: Perfectly adapts formatting and structure to the user's needs.
- 4: Good adaptation with one minor inconsistency.
- 3: Some adaptation, but formatting is inconsistent or generic.
- 2: Little to no adaptation; ignores cues.
- 1: No adaptation; formatting misaligned.

Anticipatory Flow Management

- 5: Proactively anticipates user's next steps, natural and efficient flow.
- 4: Generally anticipates needs well, with one slightly delayed transition.
- 3: Attempts to manage flow but sometimes misjudges.
- 2: Rarely anticipates; conversation feels disjointed.
- 1: Disrupts flow by consistently misinterpreting user intent.

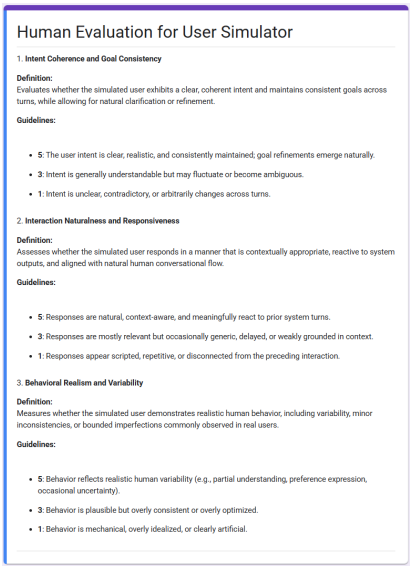
G Implementation Details

Model access and infrastructure. We conducted all experiments using a combination of Google AI Studio[1] and Vertex AI[2]. Gemini models were accessed via Google AI Studio, while all non-Gemini models were accessed through Vertex AI. To ensure broad coverage of contemporary model families and reasoning capabilities, we evaluated the following state-of-the-art models:

- `claude-sonnet-4@20250514`
- `claude-sonnet-4-5@20250929`
- `claude-haiku-4-5@20251001`
- `deepseek-ai/deepseek-r1-0528-maas`
- `gemini-2.5-flash@20250617`
- `gemini-2.5-pro@20250617`

Model roles within CoReflect. Within the CoReflect framework, different large language models were assigned fixed functional roles according to their capabilities. *Gemini 2.5 Flash* served as the backbone model for both the user simulator and the LLM-as-a-judge evaluator, where efficiency, stability, and responsiveness in multi-turn interaction and scoring were critical. *Gemini 2.5 Pro* was used as the reflective analyzer, which required stronger reasoning ability to synthesize evaluation outcomes, identify systematic behavioral patterns, and propose updates to both rubrics and conversation planning strategies.

[1] https://aistudio.google.com/
[2] https://cloud.google.com/vertex-ai



Human Evaluation for User Simulator

1. **Intent Coherence and Goal Consistency**

Definition:
Evaluates whether the simulated user exhibits a clear, coherent intent and maintains consistent goals across turns, while allowing for natural clarification or refinement.

Guidelines:

- **5**: The user intent is clear, realistic, and consistently maintained; goal refinements emerge naturally.
- **3**: Intent is generally understandable but may fluctuate or become ambiguous.
- **1**: Intent is unclear, contradictory, or arbitrarily changes across turns.

2. **Interaction Naturalness and Responsiveness**

Definition:
Assesses whether the simulated user responds in a manner that is contextually appropriate, reactive to system outputs, and aligned with natural human conversational flow.

Guidelines:

- **5**: Responses are natural, context-aware, and meaningfully react to prior system turns.
- **3**: Responses are mostly relevant but occasionally generic, delayed, or weakly grounded in context.
- **1**: Responses appear scripted, repetitive, or disconnected from the preceding interaction.

3. **Behavioral Realism and Variability**

Definition:
Measures whether the simulated user demonstrates realistic human behavior, including variability, minor inconsistencies, or bounded imperfections commonly observed in real users.

Guidelines:

- **5**: Behavior reflects realistic human variability (e.g., partial understanding, preference expression, occasional uncertainty).
- **3**: Behavior is plausible but overly consistent or overly optimized.
- **1**: Behavior is mechanical, overly idealized, or clearly artificial.

Figure 5: Human evaluation form used to assess the quality of simulated user behavior.

Co-evolution procedure. CoReflect operates through an iterative co-evolution process in which dialogue simulation, evaluation, and rubric refinement are repeatedly coupled. Across experiments, we performed three refinement iterations in total. In each iteration, the conversation planner generated structured interaction templates that guided the user simulator, whose resulting conversations were evaluated by the LLM-as-a-judge under the current rubric definitions. The reflective analyzer then aggregated evaluation signals across conversations and models to identify systematic weaknesses and emerging behaviors, which were used to update both rubric descriptions and planner templates before the next iteration.

H Human Validation of the User Simulator

To validate the quality and realism of the simulated user behavior, we conducted a human evaluation study with three volunteer annotators. Each annotator independently assessed ten randomly sampled conversations generated by the user simulator, resulting in a total of 30 ratings per evaluation aspect. Ratings were provided on a 5-point Likert scale along three dimensions: (1) *Intent Coherence and Goal Consistency*, (2) *Interaction Naturalness and Responsiveness*, and (3) *Behavioral Realism and Variability*. The evaluation form used in this study is shown in Figure 5.

Given the inherent subjectivity of human judgments, we measured inter-annotator agreement using Fleiss' Kappa (κ). The average agreement score across all metrics was $\kappa = 0.68$, corresponding to *substantial agreement*. This level of consistency indicates that the ratings reflect a shared assessment of simulator quality rather than random variation across annotators.

Evaluation Metric	Rating (Mean $\pm$ Std)	Kappa (κ)
Intent Coherence & Consistency	4.53 ± 0.63	0.72
Interaction Naturalness	4.27 ± 0.78	0.65
Behavioral Realism & Variability	4.10 ± 0.84	0.61

Table 7: Human evaluation results ($N = 30$ ratings per metric). Scores are reported on a 5-point Likert scale. Fleiss' Kappa values above 0.60 indicate substantial inter-annotator agreement.

Quantitative results are summarized in Table 7. Across all three dimensions, the simulator achieved consistently high mean scores, with strong agreement among annotators, supporting its effectiveness in producing realistic and human-like user behavior.

I Use of AI Assistants

In accordance with the ACL Publication Ethics Policy, we did not employ AI assistants to generate the initial draft or the core innovative ideas of this paper. We utilized AI tools to perform editing for improved fluency and grammatical correctness. Beyond these applications and the specific experimental uses detailed in the methodology, such as synthetic data generation and model testing, no other AI tools were used in the research process.

J Additional Experimental Results

Test Models	Task Completeness				User-Centric Personalization				Model rating
	ODI	DCA	FTP	**avg.**	AFM	OSF	SSA	**avg.**	
Claude Sonnet 4	4.70	4.75	4.54	4.66	4.73	4.75	4.76	4.75	4.71
Claude Sonnet 4.5	4.60	4.72	4.39	4.57	4.54	4.70	4.68	4.64	4.61
Claude Haiku 4.5	4.55	4.70	4.29	4.51	4.66	4.68	4.69	4.68	4.60
DeepSeek-R1	4.72	4.71	4.67	4.70	4.71	4.68	4.71	4.70	4.70
Qwen3-Next	4.74	4.51	4.70	4.65	4.71	4.75	4.75	4.74	4.70
Gemini 2.5 Pro	**4.79**	**4.79**	**4.74**	**4.77**	**4.77**	**4.79**	**4.77**	**4.78**	**4.78**
Gemini 2.5 Flash	4.76	4.74	4.54	4.68	4.72	4.73	4.76	4.74	4.71

Table 8: Model performance across all rubrics at iteration $t = 1$. (ODI: Output Delivery Integrity, DCA: Domain Conceptual Alignment, FTP: Functional Task Progression; AFM: Anticipatory Flow Management, OSF: Output Structure Fit, SSA: Sustained Style Adherence).

Test models	Task Completeness				User-Centric Personalization				Overall rating
	ODI	DCA	FTP	**avg.**	AFM	OSF	SSA	**avg.**	
Claude Sonnet 4	4.70	4.79	4.39	4.63	4.77	4.78	4.79	4.78	4.71
Claude Sonnet 4.5	4.52	4.74	3.99	4.42	4.37	4.73	4.66	4.59	4.51
Claude Haiku 4.5	4.32	4.71	3.75	4.26	4.60	4.66	4.69	4.65	4.46
DeepSeek-R1	4.72	4.71	4.63	4.69	4.71	4.66	4.71	4.69	4.69
Qwen3-Next	4.77	4.30	4.69	4.59	4.70	4.79	4.78	4.76	4.68
Gemini 2.5 Pro	**4.88**	**4.88**	**4.78**	**4.85**	**4.82**	**4.88**	**4.84**	**4.85**	**4.85**
Gemini 2.5 Flash	4.81	4.77	4.36	4.65	4.72	4.74	4.80	4.75	4.70

Table 9: Model performance across all rubrics at iteration $t = 2$. (ODI: Output Delivery Integrity, DCA: Domain Conceptual Alignment, FTP: Functional Task Progression; AFM: Anticipatory Flow Management, OSF: Output Structure Fit, SSA: Sustained Style Adherence).

Component	**Before Reflection**	**After Reflection**
Observed issue	Responses were semantically correct and complete, but later turns sometimes reverted from the requested structure, e.g., from a table to inline narrative text.	The reflective analyzer identifies structural consistency as a distinct generalization issue across heterogeneous multi-turn scenarios.
Rubric description	**Output Structure Fit**: Appropriateness and clarity of output formatting.	**Output Structure Fit**: Alignment of formatting and structural organization with explicit user preferences or previously established output schema.
Scale anchor: 5	Clear and well-organized formatting.	Perfectly adapts formatting and structure to the user's needs and maintains cross-turn consistency.
Scale anchor: 4	Mostly clear formatting with minor issues.	Good adaptation with one minor structural inconsistency.
Scale anchor: 3	Acceptable formatting but somewhat generic or inconsistent.	Some adaptation, but formatting is inconsistent or partially generic relative to the requested schema.
Scale anchor: 2	Poor formatting or weak structural organization.	Little to no adaptation; largely ignores structural cues.
Scale anchor: 1	Severely disorganized output.	No adaptation; formatting misaligned with explicit schema.

Figure 6: Details of the example illustrated in Figure 4.