Decoupled Analysis-Judging: An Automated Creativity Evaluator Using LLMs in Complex Multi-step Creativity Tasks

Xiangyu Wang[1,*], Jin Wu[2,3,*], Xiaoyu Li[4], Chanjin Zheng[2,†], Yifeng Zhou[5]
[1]Department of Educational Psychology, East China Normal University,
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University,
[3]School of Computer Science and Technology, East China Normal University
[4] School of Education and Intelligent Education Research Center, Yangzhou University
[5]School of Data Science and Engineering, East China Normal University
{51274118009, 52275901018}@stu.ecnu.edu.cn, chjzheng@dep.ecnu.edu.cn

Abstract

Automated evaluation of creativity tasks remains challenging for LLM-as-a-Judge, as LLM is susceptible to biases such as verbosity bias and leniency bias. Such limitations are particularly evident in Contextually-Grounded and Procedurally-Structured Tasks (CGPST), a complex multi-step creativity task where inter-step dependencies, highly subjectivity, and wide scoring ranges lead to more unstable and biased judgments. Existing approaches either rely on task-specific training or directly apply LLM-as-a-Judge, both of which struggle to ensure reliable evaluation under such complexity. To bridge these gaps, we propose **CreaEval**, an automated **crea**tivity **eval**uator for CGPST that decouples typical LLM-as-a-Judge into analysis and judging. Correspondingly, CreaEval involves two critical phases: **Memory-augmented Analysis**, a SoT-LLM converts multi-step responses into structured evaluation evidence, incorporating cross-step memory; and **Evidence-based Judging**, a Judge-LLM uses the extracted evidence for judging without accessing raw responses. Comprehensive experiments show that CreaEval achieves an average performance improvement of 22.74% over the second-best baselines across CGPST and two classic simple creativity tasks, demonstrating its generalizability. The code is available at https://github.com/Jaong/CreaEval.

1 Introduction

The creativity of Large Language Models (LLMs) has attracted increasing attention, with applications in creative writing, novel mathematical reasoning, and other creative domains (Kumar et al., 2025; Lin et al., 2025; Ye et al., 2025b). However, current automated evaluation methods mainly focus on simple creativity tasks, such as the Alternate Uses Task (AUT) (Lu et al., 2024) and the Torrance Test of Creative Thinking (TTCT) (Kumar et al., 2025), while the automated evaluation of complex multi-step creativity tasks remains underexplored.

Among creativity tasks, Contextually-Grounded and Procedurally-Structured tasks (**CGPST**) (Wang et al., 2026b) is particularly challenging due to its high complexity, including multi-step and scenario-based features. Each task on CGPST is grounded in a complete future scenario and needs to solve multiple interdependent steps (Treffinger, 1995; Treffinger et al., 2012b).

Existing automated evaluation methods for the above creativity tasks mainly fall into two categories. The first involves training task-specific models, which require additional training resources and annotated datasets, making them costly to develop (Do et al., 2024; Wang and Liu, 2025; Li and Pan, 2025). The other is LLM-as-a-Judge, a training-free approach that has become increasingly dominant due to the strong capabilities of LLMs (Zheng et al., 2023; Li et al., 2025; Ye et al., 2025b). However, directly applying LLM-as-a-Judge to CGPST yields low agreement, as LLMs are often susceptible to biases such as verbosity bias and leniency bias, making it difficult to accurately capture subtle score differences in subjective dimensions (Wang et al., 2026b).

Applying automated evaluation to CGPST presents several key challenges compared to other simple creativity tasks. **(1)** The task is **highly complex** compared to traditional creativity tasks. Each task requires evaluating multiple interdependent steps based on a scenarios, with each step assessed across multiple dimensions. **(2)** The task is **inherently subjective**, as there are no fixed answers and earlier responses influence subsequent ones, resulting in highly diverse outputs (Zhao et al., 2025; Wang et al., 2026b). **(3)** Some dimensions involve **large scoring ranges** (e.g., 10-level scoring), which further increases scoring instability.

To address these challenges, we propose **CreaE-**

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 17272–17308
October 24-29, 2026 ©2026 Association for Computational Linguistics

Step	Requirement	Dimensions (Score Range)	Targeted Ability
Step-1: Identify Challenges	Identify up to 8 reasonable challenges based on the future scenario.	Fluency (0-8), Flexibility (0-8), Elaboration (0-16), Originality (0-16)	Divergent Thinking
Step-2: Select an Underlying Problem	Select the most promising and meaningful challenge from Step-1 as the underlying problem.	Integrity (0-10), Focus (1-10), Adequacy (1-10)	Convergent Thinking
Step-3: Produce Solutions	Generate up to 8 solutions for the underlying problem from Step-2.	Fluency (0-8), Flexibility (0-8), Elaboration (0-16), Originality (0-16)	Divergent Thinking
Step-4: Select Criteria	Generate 5 evaluation criteria for the solutions from Step-3.	Correctly Written (0-5), Relevance (0-15)	Critical Thinking
Step-5: Apply Criteria to Top Solution	Rank the solutions from Step-3 using the criteria from Step-4 and select the highest-scoring solution.	Correctly Used (0-5)	Logical Thinking
Step-6: Develop an Action Plan	Develop the top solution from Step-5 into a comprehensive action plan to address the underlying problem from Step-2.	Relevance (1-5), Effectiveness (1-5), Criteria (1-5), Impact (1-5), Humaneness (1-5), Development (1-10)	Comprehensive Problem-Solving

Table 1: Step-wise Information on CGPST. See Appendix A for a full description.

val, an automated **crea**tivity **eval**uator using LLMs in complex multi-step creativity tasks such as CGPST. Inspired by human evaluation practices, where raters first analyze responses across each dimension before assigning rubric-based scores (Klein et al., 1998; Harsch and Martin, 2013), CreaEval decouples the evaluation process into two phases: **Memory-augmented Analysis** and **Evidence-based Judging**. In the first phase, a SoT-LLM incrementally organizes raw responses into structured intermediate evaluation evidence in the form of Structure-of-Thought (SoT) (Qi et al., 2025; Wang et al., 2026a). Considering the interdependencies among CGPST steps, CreaEval further introduces a memory mechanism to maintain cross-step coherence during analysis. In the second phase, each Judge-LLM assigns scores based on the extracted evidence and predefined rubrics, without accessing the original responses, thereby enabling evidence-grounded scoring.

Compared with typical LLM-as-a-Judge, this decoupled design constrains judging with structured evidence, thereby narrowing the range of plausible judgments. This not only improves scoring accuracy and stability, but also mitigates verbosity and leniency biases. The contributions are threefold:

- We propose **CreaEval**, an novel automated **crea**tivity **eval**uator for complex and subjective creativity tasks such as CGPST that decouples typical LLM-as-a-Judge into Memory-augmented Analysis and Evidence-based Judging.
- Extensive experiments show that CreaEval achieves an average human-LLM agreement of 0.64 (quadratic weighted kappa, QWK) on CGPST, outperforming the second-best baseline (supervised training) by 0.17.
- Further analysis reveals that the decoupled design enhances scoring stability across dimensions and mitigates verbosity and leniency biases in LLM-as-a-Judge, providing new insights into evaluating creativity tasks.

2 Related Work

2.1 Creativity Tasks

Traditional creativity tasks originate from educational and psychological studies, such as AUT, which requires participants to generate as many novel uses as possible for a common object (Lu et al., 2024; Zhao et al., 2025; Organisciak et al., 2023; Hadas and Hershkovitz, 2024), and TTCT, which assesses creativity through responses to open-ended and unconventional scenarios (Torrance, 1966; Kumar et al., 2025). However, these tasks are typically single-step and structurally simple, limiting their ability to capture complex creative processes. Recently, Wang et al. (2026b) proposed CGPST, a multi-step, scenario-based benchmark for evaluating creative problem-solving abilities of LLMs. As shown in Table 1, each task is grounded in a complete future scenario and requires LLMs to sequentially solve six interdependent steps, involving diverse abilities (Treffinger, 1995; Treffinger et al., 2012b). Compared to traditional creativity tasks, CGPST exhibits significantly higher complexity.

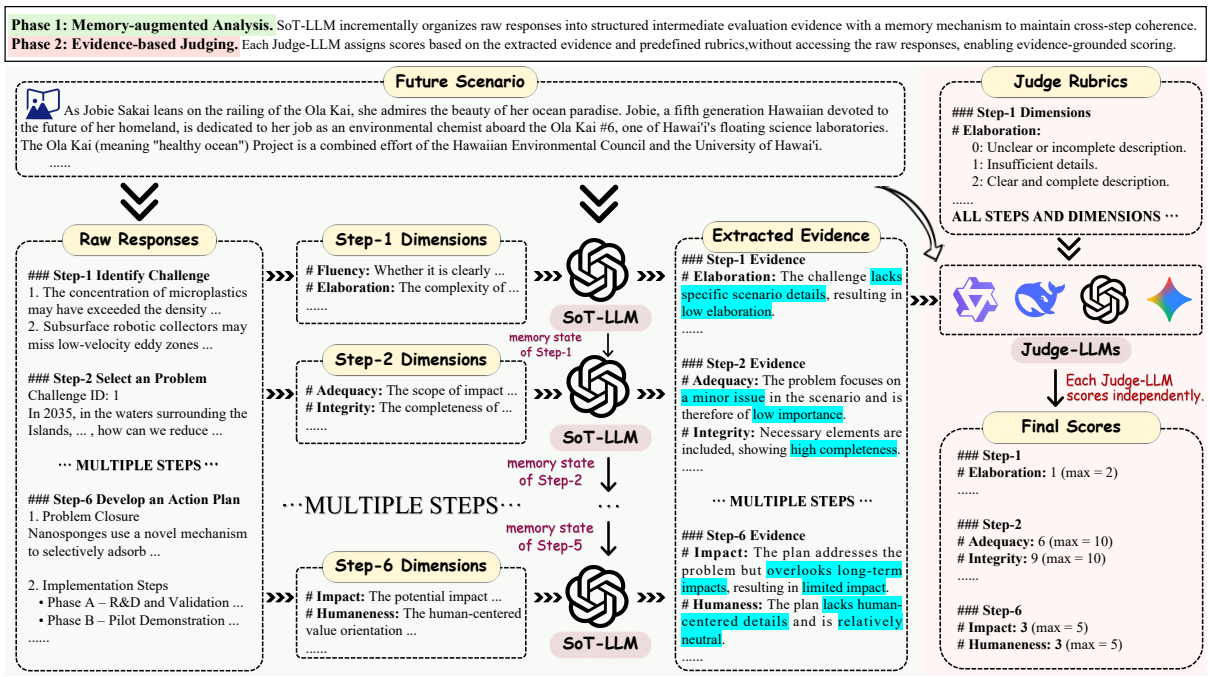

Figure 1: Our proposed CreaEval framework. CreaEval decouples the evaluation process into two phases: **1) Memory-augmented Analysis.** SoT-LLM first extracts intermediate evaluation evidence (blue) based on the responses and dimensions in a step-by-step manner. **2) Evidence-based Judging.** Each Judge-LLM then performs scoring based on the extracted evidence and predifined rubrics without accessing raw responses.

2.2 LLM-as-a-Judge for Creativity Evaluation

Recently, several studies have explored automated evaluation methods for creativity tasks (Zheng et al., 2023; Liang et al., 2024; Li et al., 2025; Ye et al., 2025b). For example, Zhao et al. (2025) leverages GPT-4 to generate TTCT-inspired datasets and employs LLMs for scoring responses. Lu et al. (2024) applies LLM-as-a-Judge to AUT tasks, achieving an average human–LLM agreement of 0.49 (Kendall's τ) across four dimensions, even surpassing inter-human agreement (0.39), suggesting the potential of LLM-as-a-Judge to improve scoring reliability in creativity tasks. However, existing studies mainly focus on relatively simple and traditional creativity benchmarks. Directly applying LLM-as-a-Judge to more complex tasks such as CGPST remains challenging. For instance, Wang et al. (2026b) shows that a direct LLM-as-a-Judge approach with few-shot prompting achieves only 0.31 (pearson correlation coefficient, PCC) human–LLM agreement, highlighting the need for more reliable and effective evaluation frameworks for complex multi-step creativity tasks.

3 Methodology

Existing LLM-as-a-Judge fail to decouple analysis from judging, instead directly mapping from raw responses $R^s = \{R^s_t\}_{t=1}^{N}$ (N=6) to a score list S. This leads to the LLM's inability to effectively capture key intermediate evidence when handling complex creativity tasks. To address this limitation, we propose **CreaEval**, a novel evaluation framework that explicitly models the evaluation process in a structured manner, decoupling analysis and judging. As illustrated in Figure 1, CreaEval consists of two sequential phases: (1) Memory-augmented Analysis; and (2) Evidence-based Judging.

3.1 Phase 1: Memory-augmented Analysis

Given the raw response R^s_t and the corresponding dimension description D_t in Step-t, SoT-LLM performs structured evidence extraction in a step-by-step manner. Specifically, it iteratively processes each step of the CGPST by jointly considering the task scenario *Scenario*, the step response R^s_t, and the evaluation dimensions D_t. Within Step-t, SoT-LLM in CreaEval extracts multi-dimensional evidence represented as a structured mapping $e_t = \{\langle k_i, v_i \rangle\}_{i=1}^{|\mathcal{D}_t|}$, where each pair $\langle k_i, v_i \rangle$ denotes an evaluation dimension k_i and its corresponding ev-

idence v_i. This progressive design prevents SoT-LLM from being overwhelmed by all subjective information at once, thereby improving the reliability and consistency of the extracted evidence.

The steps of CGPST exhibit strong temporal dependencies, as responses in later steps are conditioned on decisions made in earlier ones. For instance, the solution proposed in Step-3 is designed to address the problem identified in Step-2. To model these dependencies, CreaEval introduces a memory mechanism that preserves step-relevant memory state m_t. Specifically, in addition to generating evidence e_t, SoT-LLM produces a corresponding $m_t = Summary(R_t^s, M_{t-1})$, which summarizes the key information from R_t^s and is passed as contextual input to subsequent steps as shown in Figure 1. This mechanism retains cross-step contextual dependencies, thereby augmenting the evidence extraction process and improving accuracy, defined as:

$$(e_t, m_t) = f_{SoT}(Scenario, R_t^s, D_t, M_{t-1}) \quad (1)$$

where e_t and m_t denote the evaluation evidence and memory state extracted by SoT-LLM in Step-t, respectively. $D = \{D_t\}_{t=1}^N$ and $R^s = \{R_t^s\}_{t=1}^N$ denote the dimension information and raw responses across all steps, respectively. $M_{t-1} = \{m_j\}_{j=1}^{t-1}$ represents the accumulated memory of previous steps to preserve cross-step dependency.

3.2 Phase 2: Evidence-based Judging

In this phase, Judge-LLM performs scoring grounded in the structured evidence generated during Phase 1. Rather than accessing the raw responses R^s directly, Judge-LLM takes the aggregated evidence E , the scoring rubrics R^u, and *Scenario* as input, and generates multi-step and multi-dimensional scores S over the entire response in a single pass, defined as:

$$S = f_{Judge}(Scenario, E, D, R^u) \quad (2)$$

where $E = \{e_t\}_{t=1}^N$ and $R^u = \{R_t^u\}_{t=1}^N$ denote the evidence extracted by SoT-LLM and the scoring rubrics for all steps, respectively.

As illustrated in Figure 1, consider the *Adequacy* dimension in Step-2: the extracted evidence (e.g., "*low importance*" and "*a minor issue*") indicates weak performance along this dimension, upon which Judge-LLM assigns a relatively low score of 6 out of 10. This evidence-grounded judging enhances both the accuracy and stability of the final scoring.

4 Experiment

4.1 Challenging Dataset

We conduct experiments on the Contextually-Grounded and Procedurally-Structured Tasks (CGPST) dataset (Wang et al., 2026b), a complex multi-step creativity benchmark. It possesses a strong psychological foundation based on the Future Problem Solving Program International (FPSPI) (Treffinger et al., 2012a; Alt et al., 2022; Wang et al., 2026b), an active international creativity competition framework with over 50 years of history founded by psychologist Ellis Paul Torrance (Torrance, 1966). Each task on CGPST is grounded in a future scenario and needs to complete six interdependent steps, each associated with multiple scoring dimensions, as shown in Table 1. The dataset contains 10 different scenarios, each with 20 samples, resulting in a total of 200 samples. Each sample includes complete responses of six steps and is annotated with calibrated scores from two human evaluators. The inter-rater reliability between the two human evaluators reaches 0.84, indicating that the CGPST dataset is of high quality. Due to its strong subjectivity, cross-step dependencies, fine-grained multi-dimensional scoring, and wide score ranges, CGPST poses significant challenges for automated evaluation. More details are shown in Appendix A.

To further demonstrate the generalizability of CreaEval beyond complex multi-step creativity task, we also evaluate our method on two classic creativity benchmarks with different task structures: (1) **Alternative Uses Task (AUT)** (Organisciak et al., 2023; Hadas and Hershkovitz, 2024), a widely used creativity test requiring participants to generate novel uses for a given object, scored exclusively on *Originality*. (2) **Torrance Test of Creative Writing (TTCW)** (Chakrabarty et al., 2024), a narrative generation task requiring creative story writing based on a given plot, evaluated across four dimensions (*Fluency*, *Flexibility*, *Originality*, *Elaboration*). While we have incorporated AUT and TTCW dataset to ensure generalizability, CGPST remains our primary benchmark due to its comprehensive assessment and structural complexity.

4.2 Baselines

We compare our CreaEval with the following baselines. Detailed implementation details of all baselines are provided in Appendix B.

Method	Step-1				Step-2			Step-3			
	Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	Fluency	Flexibility	Elaboration	Originality
Direct Score	0.5577	0.1771	0.1074	0.3552	0.184	0.042	0.0339	0.7138	0.2965	0.3177	0.3161
CoT	0.4906	0.1779	0.1298	0.2974	0.1729	0.027	0.038	0.7182	0.2747	0.324	0.3193
ToT	0.5311	0.1745	0.1269	0.3149	0.1554	0.0573	0.0349	0.7147	0.3013	0.315	0.2983
GoT	0.5244	0.1775	0.1281	0.3364	0.1656	0.022	0.0013	0.7464	0.3189	0.2909	0.3251
TaT	0.4766	0.1853	0.1336	0.2894	0.1954	0.0384	0.0171	0.6365	0.3026	0.3307	0.264
SaMer	0.1589	0.2668	0.2821	0.276	0.204	0.1342	0.107	0.2117	0.1698	0.2112	0.2461
SFT	0.5027	0.4324	0.4765	0.5353	0.476	0.4869	0.4194	0.4723	0.4662	0.4756	0.4541
CreaEval	**0.7505***	**0.6665***	**0.656***	**0.665***	**0.7756***	**0.554***	**0.4831**	**0.8315***	**0.6975***	**0.6969***	**0.6965***

Method	Step-4		Step-5	Step-6						AVG
	Correctly Written	Relevance	Correctly Used	Relevance	Effectiveness	Criteria	Impact	Humaneness	Development	
Direct Score	0.4832	0.0798	0.4627	0.0744	0.1217	0.1491	0.1104	0.0756	0.1714	0.2415
CoT	0.3723	0.1018	0.2836	0.0328	0.1094	0.1641	0.1038	0.0548	0.1825	0.2187
ToT	0.379	0.1	0.3128	0.0415	0.1031	0.1634	0.0975	0.0542	0.1816	0.2229
GoT	0.4551	0.0438	0.0828	0.0201	0.1077	0.183	0.1137	0.0559	0.1821	0.214
TaT	0.3191	0.0679	0.3713	0.0324	0.1108	0.1279	0.0735	0.0419	0.15	0.2082
SaMer	0.1661	0.1238	0.3594	0.1829	0.2044	0.1749	0.1583	0.1326	0.1052	0.1938
SFT	0.4599	0.4572	0.5317	0.4408	0.4469	0.4582	0.4694	0.4065	0.4617	0.4665
CreaEval	**0.7933***	**0.535***	**0.9439***	**0.5366**	**0.5214***	**0.5276***	**0.4952**	**0.4388**	**0.5119**	**0.6388***

Table 2: Consistency (QWK) results across different methods and dimensions on the CGPST benchmark. **Bold** indicates the highest score and underline indicates the second highest score. The AVG column summarizes the overall average score. The asterisk (*) marks statistically significant improvements ($p < 0.05$, t-test) over the second-best method.

- **Direct Score**: This method directly assigns a score to a complete response without intermediate steps or decomposition.
- **Chain-of-Thought** (CoT) (Wei et al., 2022): This method decomposes the problem into intermediate steps and solve each before giving the final answer.
- **Tree-of-Thought** (ToT) (Yao et al., 2023): This method actively maintains a tree of thoughts, where each thought is a coherent language sequence that serves as an intermediate step toward problem solving.
- **Graph-of-Thought** (GoT) (Besta et al., 2024): This method models the problem-solving process as a graph with more flexible thought transformations compared to ToT.
- **Table as Thought** (TaT) (Sun et al., 2025): This method uses a table to represent structured thoughts. Although this method shares the same structured representation, it does not decouple analysis and judging, making it an **undecoupled variant of CreaEval**.
- **Supervised Fine-Tuning** (SFT): We fine-tunes `Qwen3.5-9B`[1] on a 7:3 train-test split of the dataset, and reports results on test set.
- **SaMer** (Feng et al., 2025): SaMer is a scenario-aware multi-dimensional LLM evaluator that adaptively identifies and weights dimensions according to different scenarios.

Unlike CGPST, both AUT and TTCW are single-turn generation tasks without procedural process and dependencies. Consequently, we omit structure-based baselines like ToT and GoT for these benchmarks. Additionally, we additionally incorporate a reference-based approach (Li et al., 2025) using human-written stories as evaluation references as a train-free baseline for TTCW.

4.3 Experimental Setup

We employ four LLMs (`qwen3.6-plus`, `gpt-5.4`, `deepseek-v4-pro`, `gemini-3.1-pro`) as Judge-LLMs for all methods. We use `qwen3.6-plus` as SoT-LLM for structured evidence extraction. Full LLM details are provided in Appendix E. We report three agreement metrics: Pearson Correlation Coefficient (PCC), Quadratic Weighted Kappa (QWK), and Intra-class Correlation Coefficient (ICC), with details in Appendix C. Additionally, we set the temperature of all LLMs to 0.2. Appendix F shows a pilot study on temperature selection. Appendix D presents the full prompts.

[1] https://huggingface.co/Qwen/Qwen3.5-9B

Method	AUT	TTCW				
	Originality	Fluency	Flexibility	Originality	Elaboration	AVG
Direct Score	0.2731	0.5367	0.4234	0.2508	0.2145	0.3558
CoT	0.219	0.3392	0.1879	0.2227	0.2744	0.256
SFT	0.6924	0.3888	0.2783	0.2252	0.2095	0.2755
Reference-based	–	0.6484	0.5775	0.5802	0.5174	0.581
CreaEval	**0.7382**	**0.7458***	**0.8052***	**0.6249**	**0.652***	**0.707***

Table 3: Consistency (QWK) results across different methods and dimensions on the AUT and TTCW benchmark. **Bold** indicates the highest score and underline indicates the second highest score. The AVG column summarizes the overall average score. The asterisk (*) marks statistically significant improvements ($p < 0.05$, t-test) over the second-best method.

5 Results

5.1 Main Results

As shown in Table 2 (reporting QWK. PCC and ICC show similar trends, see Appendix H for full results), CreaEval consistently achieves the best performance (0.64), significantly outperforming all baselines on the CGPST. This indicates that CreaEval better aligns with human judgments with decoupled analysis and judging. In contrast, other training-free methods show relatively weak and similar performance (0.2-0.25), suggesting that simply increasing reasoning steps does not improve accuracy for subjective dimensions evaluation. For training-based methods, SFT achieves the second-best performance (0.47), indicating that supervised data can effectively align LLM judging with human preferences to a certain extent, but still falls short of CreaEval. Furthermore, as presented in Table 3, CreaEval also achieves superior performance on simple creativity tasks, demonstrating its strong generalizability to simpler creativity tasks beyond complex multi-step evaluations.

Performance varies across dimensions. As shown in Table 2, for dimensions with large range such as Step-2 (*Focus*, *Adequacy*) and Step-4 (*Relevance*), most training-free baselines remain below 0.2, indicating that they struggle to accurately distinguish score differences within wide ranges. In contrast, Step-5 is relatively less subjective, as it mainly involves verifying whether scoring vectors satisfy a non-repetitive ranking structure. Nevertheless, most training-free methods still achieve only weak performance (around 0.08–0.46 v.s. CreaEval 0.94), as they retain redundant information from previous steps without effective evidence extraction, which further disrupts subsequent evaluations. Step-6 is the most challenging step, requiring holistic integration across all previous steps to produce a final action plan. In Step-6, all training-free baselines collapse to very weak performance (<0.2), whereas CreaEval still maintains moderate consistency (around 0.5). This further validates the effectiveness of the decoupled analysis-judging design.

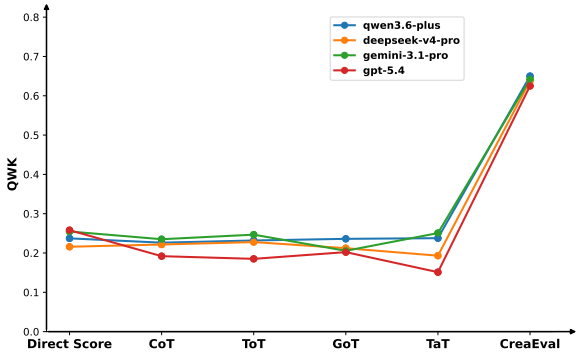

Figure 2: Judge QWK Consistency Across Methods.

Furthermore, as shown in Figure 2, we analyze the consistency of four Judge-LLMs under each method. CreaEval exhibits nearly identical performance across all judges, indicating stable and robust evaluation behavior. This stability stems from replacing raw responses with structured evidence, which reduces the impact of subjective variation in the raw responses in scoring, further demonstrating that CreaEval is a model-agnostic framework that does not rely on the specific Judge-LLM. In contrast, training-free methods show noticeable fluctuations across different judges due to variations in model capability, resulting in more diverse scoring behavior. We further provide an in-depth analysis of scoring stability in Section 6.3.

5.2 Ablation Study

We conduct ablation studies to evaluate the effectiveness of key components in CreaEval. Specifically, we consider four variants: (1) **w/o Memory**, where the memory mechanism in Phase 1 is re-

Method	PCC	ICC	QWK
CreaEval	**0.69**	**0.64**	**0.64**
w/o Memory	0.57	0.49	0.49
w/o Evidence	0.33	0.24	0.24
w/o Decoupled	0.29	0.2	0.21
w/o Step-wise	-0.01	0.01	0.01

Table 4: Ablation results.

moved, and each step is processed independently without cross-step dependency; (2) **w/o Evidence**, where Judge-LLM directly performs scoring based on the raw responses without extracting structured evidence, corresponding to Direct Score baseline; (3) **w/o Decoupled**, where analysis and judging are performed jointly within a single LLM without separation, corresponding to TaT baseline; and (4) **w/o Step-wise**, where evidence extraction is performed in a single pass over all six steps instead of step-by-step iterative extraction.

Table 4 presents the results. Removing memory leads to performance degradation (QWK = 0.49, 23%↓), demonstrating that the memory mechanism is necessary for maintaining cross-step coherence. Without evidence, the performance decreases (QWK = 0.24, 63%↓), indicating that evidence-based scoring is more accurate than direct scoring. Similarly, when analysis and judging are not decoupled, performance further degrades (QWK = 0.21, 67%↓), suggesting that coupling the two processes introduces analytical bias into scoring. Meanwhile, extracting evidence for all steps at once leads to a collapse in performance (QWK = 0.01, 98%↓), indicating that excessive input information significantly harms SoT-LLM's evidence extraction accuracy. For example, SoT-LLM provides evidence for some dimensions but only produces binary judgments (e.g., *"Yes"* or *"No"*) for others, which leads to significant degradation in scoring quality.

6 Discussion

In this section, we conduct a comprehensive analysis of CreaEval, including robustness, bias mitigation, scoring stability, efficiency, and case study, further validating the effectiveness of CreaEval beyond simple improvements in scoring accuracy.

6.1 Impact of SoT-LLM in CreaEval

The quality of extracted evidence directly affects the accuracy of the final scoring, making SoT-LLM a critical component. To evaluate the robustness of CreaEval, we replace SoT-LLM with other LLMs and report QWK across different LLMs combinations. As shown in Figure 4, the agreement remains consistently above 0.6 across all SoT-LLMs and Judge-LLMs combinations, indicating that CreaEval does not rely on a specific LLM and can serve as a general and robust evaluation framework. PCC and ICC results are provided in Appendix I.

6.2 Bias Mitigation in CreaEval

In LLM-as-a-Judge, LLMs are prone to various evaluation biases during scoring. In automated evaluation of creative tasks, two biases are particularly critical: **(1) leniency bias**, where LLMs tend to assign overly high scores to subjective dimensions due to their inherent sycophantic tendencies (Ye et al., 2025a; Gupta et al., 2026); and **(2) verbosity bias**, where LLM favor longer responses over shorter ones, even when the latter are clearer or of higher quality (Zheng et al., 2023).

For leniency bias, we analyze the heatmap between different methods and human scores on *Originality* of Step-3, as shown in Figure 3. The results indicate that all training-free methods exhibit a clear inclination to cluster their scores within the high-value region. In contrast, CreaEval significantly mitigates this tendency, demonstrating a more balanced score distribution that aligns closely with human annotations. Heatmaps for additional dimensions are provided in Appendix J.

For verbosity bias, we examine PCC between response length of Step-6 and *Development* (the level of detail of action plan) scores to measure their linear correlation. Figure 5 illustrates the relative changes for training-free methods against Direct Score. Only CreaEval shows a decrease (-5.81%), indicating that response length is less correlated with scores compared to Direct Score and thus verbosity bias is mitigated. However, CoT, GoT, and ToT even exacerbate the bias despite introducing more complex reasoning paths. TaT shows a slight increase (+0.82%), suggesting that incorporating analysis can partially reduce verbosity bias. However, due to its coupled analysis-and-judging design, its effectiveness remains inferior to CreaEval.

6.3 Scoring Stability Analysis

We analyze scoring stability by measuring the distribution of inter-Judge variance for each method. Specifically, for each sample and dimension, we compute the variance of scores assigned by four

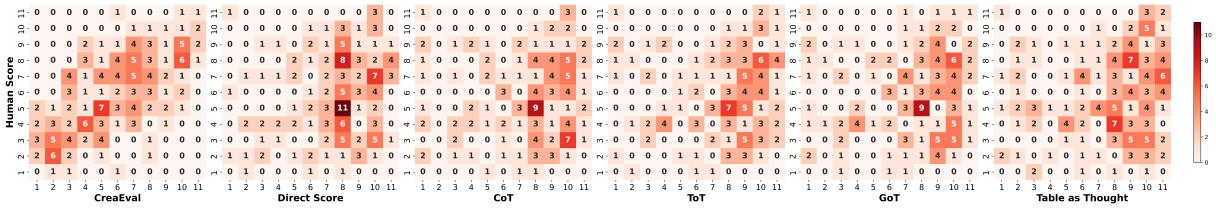

Figure 3: Headmap between human score and all methods on *Originality* of Step-3.

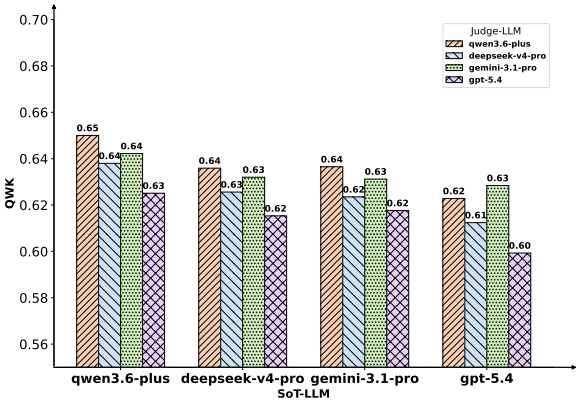

Figure 4: QWK results across different SoT-LLMs.

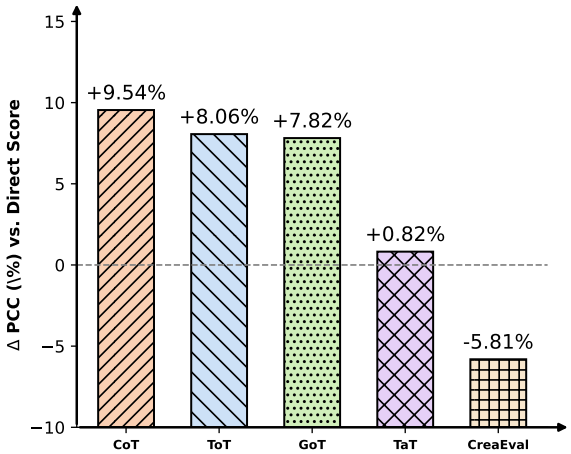

Figure 5: Relative changes in PCC against Direct Score. PCC measures the linear correlation between Step-6 response length and *Development* scores.

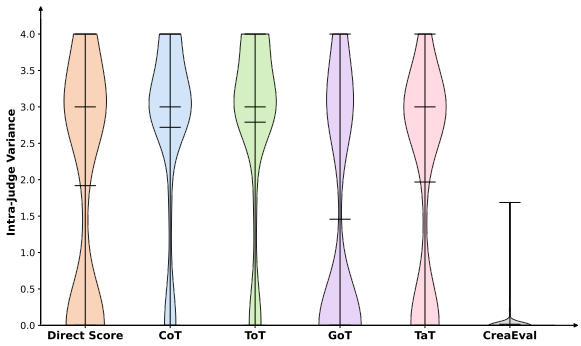

Figure 6: Distribution of inter-Judge variance across the entire dataset for *Correctly Used* in Step-5.

Judge-LLMs. This yields 200 variance values per method, corresponding to all samples on CGPST dataset, which are then analyzed as a distribution. Figure 6 illustrates the variance distribution for *Correctly Used* in Step-5. CreaEval exhibits consistently lower inter-Judge variance, indicating the four Judge-LLMs produce highly similar scores, leading to higher scoring stability. This is attributed to its evidence-based judging design, where all Judges rely on shared extracted evidence rather than raw subjective responses, leading to more consistent evaluations. In contrast, other methods show substantially higher and more dispersed variance, suggesting unstable judgments across different Judges even for the same response. Violin plots for other dimensions are provided in Appendix K.

6.4 Inference Cost Analysis

We further analyze the inference cost of different methods in terms of time and token consumption, as shown in Table 5. Overall, methods with more complex reasoning exhibit significantly higher consumption. In particular, GoT and ToT introduce substantial overhead due to expanded reasoning paths. In contrast, CoT and TaT achieves relatively low consumption. CreaEval maintains a better balance, achieving substantially lower cost than ToT and GoT while maintaining strong performance.

6.5 Case Study

As shown in Figure 7, we present a case study of Step-2. TaT, an undecoupled variant of CreaEval, tends to produce overly positive evaluations (e.g., "*a high level of focus*"), exhibiting clear leniency bias. In contrast, CreaEval decouples evidence extraction from judging, grounding scoring in structured intermediate evidence rather than raw responses. This design mitigates over-optimistic scoring (e.g., "*less focused problem*"), leading to more accurate evaluations. Case studies for other steps are provided in Appendix L.

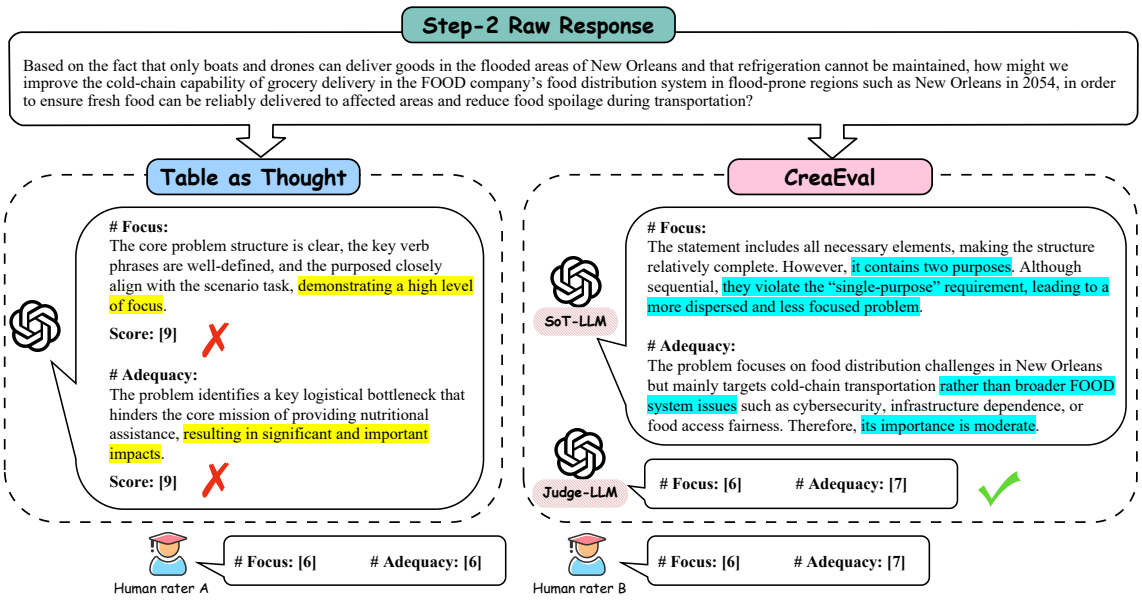

Figure 7: Case study of Step-2 comparing Table as Thought (TaT) and CreaEval. The example illustrates that TaT tends to produce overly positive and less discriminative judgments (yellow) due to coupled analysis and scoring within a single LLM, leading to leniency bias. In contrast, CreaEval separates evidence extraction and judging into distinct LLMs, enabling more grounded and constraint-aware judgments (blue).

	Time (s)		Token ($\times 10^3$)	
Method	p50	p95	p50	p95
Direct Score	30	37	13.20	15.70
CoT	54	158	66.09	73.24
ToT	200	332	269.65	297.22
GoT	479	714	567.37	620.98
TaT	72	104	16.13	18.80
CreaEval	173	249	42.85	47.13

Table 5: Time and token consumption. Both are reported in terms of the 50th and 95th percentile (p50/p95).

7 Conclusions

In this work, we propose CreaEval, an automated creativity evaluation framework for CGPST that decouples evaluation into memory-augmented analysis and evidence-based judging. SoT-LLM first converts multi-step responses into structured evidence with cross-step memory, and Judge-LLM performs scoring based solely on this evidence without accessing the raw responses. Experiments across four LLMs show that CreaEval achieves strong alignment with human judgments, consistently outperforming all baselines. Further analysis demonstrates that the decoupled design improves scoring stability and reduces verbosity and leniency biases in LLM-as-a-Judge, offering new insights into evaluating subjective creativity tasks.

Limitations

Although our work demonstrates strong effectiveness and achieves promising results, it still has several limitations. In Phase 1, we use a simple rule-based memory mechanism to maintain cross-step coherence due to the fixed step dependencies on CGPST. More advanced memory modules such as hierarchical memory modules could be explored for more general settings. In addition, to the best of our knowledge, CGPST is currently the only publicly available multi-step creativity benchmark dataset, so our experiments are conducted exclusively on this multi-step dataset. Future work will evaluate the effectiveness of CreaEval on additional datasets once they become available.

References

Dorit Alt, Yoav Kapshuk, and Heli Dekel. 2022. Promoting perceived creativity and innovative behavior: Benefits of future problem-solving programs for higher education students. *Thinking Skills and Creativity*, 47:101201.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large

language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Heejin Do, Yunsu Kim, and Gary Lee. 2024. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Kehua Feng, Keyan Ding, Jing Yu, Yiwen Qu, Zhiwen Chen, chengfei lv, Gang Yu, Qiang Zhang, and Huajun Chen. 2025. Samer: A scenario-aware multi-dimensional evaluator for large language models. In *International Conference on Learning Representations*, volume 2025, pages 40346–40367.

Manan Gupta, Inderjeet Nair, Lu Wang, and Dhruv Kumar. 2026. Context over content: Exposing evaluation faking in automated judges. *Preprint*, arXiv:2604.15224.

Eran Hadas and Arnon Hershkovitz. 2024. Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity*, 52:101549.

Claudia Harsch and Guido Martin. 2013. Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20(3):281–307.

Stephen P Klein, Brian M Stecher, Richard J Shavelson, Daniel McCaffrey, Tor Ormseth, Robert M Bell, Kathy Comfort, and Abdul R Othman. 1998. Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11(2):121–137.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025. Automated creativity evaluation for large language models: A reference-based approach. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21475–21488, Suzhou, China. Association for Computational Linguistics.

Xia Li and Wenjing Pan. 2025. Kaes: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24476–24484.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. 2025. Creativity in LLM-based multi-agent systems: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27584–27607, Suzhou, China. Association for Computational Linguistics.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *Preprint*, arXiv:2405.06373.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rui Qi, Zhibo Man, Yufeng Chen, Fengran Mo, Jinan Xu, and Kaiyu Huang. 2025. SoT: Structured-of-thought prompting guides multilingual reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11024–11039, Suzhou, China. Association for Computational Linguistics.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

Zhenjie Sun, Naihao Deng, Haofei Yu, and Jiaxuan You. 2025. Tables as thought: Exploring structured thoughts in LLM reasoning. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 19–33, Vienna, Austria. Association for Computational Linguistics.

E Paul Torrance. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Donald Treffinger, Marianne Solomon, and Deb Woythal. 2012a. Four decades of creative vision: Insights from an evaluation of the future problem solving program international (fpspi). *The Journal of Creative Behavior*, 46.

Donald J. Treffinger. 1995. Creative problem solving: Overview and educational implications. *Educational Psychology Review*, 7(3):301–312.

Donald J Treffinger, Marianne Solomon, and Deb Woythal. 2012b. Four decades of creative vision: Insights from an evaluation of the future problem solving program international (fpspi). *The Journal of Creative Behavior*, 46(3):209–219.

Jiong Wang and Jie Liu. 2025. T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236, Abu Dhabi, UAE. Association for Computational Linguistics.

Qinsi Wang, Hancheng Ye, Jinhee Kim, Jinghan Ke, Yifei Wang, Martin Kuo, Zishan Shao, Dongting Li, Yueqian Lin, Ting Jiang, Chiyue Wei, Qi Qian, Wei Wen, Helen Li, and Yiran Chen. 2026a. T2s-bench & structure-of-thought: Benchmarking and prompting comprehensive text-to-structure reasoning. *Preprint*, arXiv:2603.03790.

Xiangyu Wang, Jin Wu, Haoran Shi, Wei Xia, Jiarui Yu, and Chanjin Zheng. 2026b. Teamllm: A human-like team-oriented collaboration framework for multi-step contextualized tasks. *Preprint*, arXiv:2604.06765.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

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. 2025a. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. 2025b. Assessing the creativity of llms in proposing novel solutions to mathematical problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25687–25696.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436.

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.

A CGPST Benchmark Details

Contextually-Grounded and Procedurally-Structured tasks (CGPST) (Wang et al., 2026b) is a benchmark designed to evaluate the comprehensive problem-solving abilities of LLMs through contextualized multi-step tasks. Compared with traditional creativity benchmarks such as AUT and TTCT, CGPST is substantially more complex and therefore more challenging for automated evaluation. Specifically, CGPST exhibits four key characteristics: **Contextual Grounding**, where each task is based on a complete future scenario with rich contextual information and detailed problem settings; **Procedural Structure**, where each task consists of six interdependent and sequential steps progressively leading to the resolution of a real-world problem; **Process-Oriented Evaluation**, where all intermediate steps are systematically evaluated rather than focusing only on the final response; and **Multi-Dimensional Assessment**, where each step is evaluated across multiple comprehensive dimensions (Treffinger, 1995; Treffinger et al., 2012b; Wang et al., 2026b).

The CGPST dataset contains 10 different scenarios, each consisting of 20 complete six-step response samples, resulting in a total of 200 samples. Each sample is annotated with validated scores from two human experts. To facilitate understanding, we provide a complete future scenario in *Ocean Soup Future Scenario* box (Page 14) and an example sample in *A Complete CGPST Sample* box (Page 15). Table 6 presents the themes of all scenarios.

As shown in Table 1, CGPST consists of six sequential steps. Given a future scenario, LLMs are first required to identify up to eight challenges (**Step-1**), then select the most promising challenge as an underlying problem (**Step-2**), and generate up to eight solutions for this problem (**Step-3**). Subsequently, LLMs create five evaluation criteria for the problem and proposed solutions (**Step-4**), which are then used to rank the solutions and select the best one (**Step-5**). Finally, the selected solution is further developed into a comprehensive action plan addressing the underlying problem and generating positive impacts on the future scenario (**Step-6**). The strong interdependency across all six steps further increases the difficulty of automated evaluation. We provide detailed descriptions of dimensions for each step in Table 14 to 19.

Scenario ID	Theme
FS1	Autonomous Transportation
FS2	Terraforming
FS3	Sustainable Development
FS4	Throw Away Society
FS5	Antibiotic Resistance
FS6	Neurotechnology
FS7	Criminal Justice Systems
FS8	Biosecurity
FS9	Food Safety
FS10	Ocean Soup

Table 6: Overview of the 10 future scenarios on the CGPST dataset. Each scenario contains 20 six-step response samples, resulting in a total of 200 samples annotated with two validated human expert scores.

B Baseline Implementation Details

We provide the detailed implementation of the baseline methods as follows.

Direct Score LLM directly assigns scores to all steps of a response in a single pass. The input includes the future scenario, the full response, dimension descriptions, and corresponding rubrics.

Chain-of-Thought (CoT) Given the inherent multi-step structure of CGPST, the evaluation is performed step by step, where each step is scored sequentially. Specifically, a single response requires six sequential passes corresponding to the six steps (Wei et al., 2022).

Tree-of-Thought (ToT) ToT follows a tree-structured evaluation process, with each layer aligned to a step on CGPST. At each layer, three independent Score LLMs generate three candidate scores, forming three branches of this layer. A separate Judge LLM then selects the most appropriate score among these candidates as the final output for this step (Yao et al., 2023).

Graph-of-Thought (GoT) Built upon ToT, GoT introduces an additional refinement operation for thought transformation. Specifically, for each step, three Score LLMs first generate candidate scores, which are then refined through a self-refinement process before final selection. The Judge LLM then selects the final score from the refined candidates for each step. This refinement mechanism allows iterative improvement of intermediate scoring nodes, forming a graph-structured reasoning

process (Besta et al., 2024).

Table as Thought This method organizes reasoning within a tabular schema. In the original paper, the table is stored in JSON format, which is consistent with the representation format used in CreaEval. However, unlike CreaEval, both analysis and judging are performed by a single LLM within this framework, making it an **undecoupled variant of CreaEval**. We therefore regard it as a baseline that shares the same structured representation but does not decouple the evaluation process.

Supervised Fine-Tuning (SFT) We split the CGPST dataset into training and test sets with a ratio of 7:3. Subsequently, we perform SFT on `Qwen3.5-9B` using the instruction-response pairs constructed from the training set. After training, the model's capability is evaluated on the test set.

Scenario-aware Multi-dimensional Evaluator (SaMer) SaMer introduces a three-branch scoring head atop a frozen `Qwen3.5-9B` to enable scenario-aware, preference-driven evaluation. During training, the base LLM is frozen and only the head is optimized through a multi-objective loss combining three signals: a dimension prediction loss that learns step-to-dimension relevance, a dimension-level ranking loss that aligns pairwise preferences on individual dimensions, and an overall preference loss that learns the final winner from pairwise comparisons. At inference, only dimensions predicted as relevant receive non-zero weights via softmax, and the overall score is computed as the weighted sum of dimension scores.

C Agreement Metric Details

On the CGPST dataset, each sample is annotated by two human raters. Therefore, when computing agreement between LLM scores and human annotations, **we calculate the consistency separately with Human A and Human B for each dimension, and then average the two results as the final agreement score**. Importantly, we do not compute agreement against the averaged human score. This is to avoid misleading cases where Human A gives a score of 4 and Human B gives 8; an LLM score of 6 would appear perfectly accurate if compared to the mean of two humans (6), even though it does not truly match either annotator.

We present three metrics below to facilitate a better understanding of the experimental results.

Pearson correlation coefficient (PCC) PCC measures the **linear correlation** between predicted scores and human annotations. It evaluates how well the model predictions align with human judgments in terms of overall trend consistency.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

where x_i and y_i denote the predicted and human scores for sample i, and $\bar{x}$ and $\bar{y}$ denote their respective means.

Quadratic weighted kappa (QWK) QWK measures the agreement between two raters while taking into account the **ordinal nature** of the ratings and the degree of disagreement. It is particularly suitable for discrete or ordinal scoring tasks and is one of the most commonly used metrics in automated scoring (Ridley et al., 2021; Do et al., 2024).

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{4}$$

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{5}$$

where O_{ij} and E_{ij} are the observed and expected agreement matrices, respectively, and w_{ij} is the quadratic weight.

Intra-class correlation coefficient (ICC) ICC is widely used to assess inter-rater reliability by measuring the proportion of variance attributable to differences between subjects relative to total variance. It reflects the consistency of quantitative measurements across different raters.

$$ICC = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} \tag{6}$$

where $\sigma^2_{\text{between}}$ and σ^2_{within} denote the between-subject and within-subject variance, respectively.

Ocean Soup Future Scenario

As Jobie Sakai leans on the railing of the Ola Kai, she admires the beauty of her ocean paradise. Jobie, a fifth generation Hawaiian devoted to the future of her homeland, is dedicated to her job as an environmental chemist aboard the Ola Kai #6, one of Hawai'i's floating science laboratories. The Ola Kai (meaning "healthy ocean") Project is a combined effort of the Hawaiian Environmental Council and the University of Hawai'i.

Now, in 2035 after 15 active years, the project is struggling to live up to its nickname: the OK Project. Originally, the OK Project focused on the waters affected by Hawai'i's island-generated pollutants. Public interest in the project led to a resurgence in eco-education; recycling and the reduced use of plastics became an accepted part of island life for Hawai'i residents. The "adopt a beach" clean-up program became a popular draw for eco-tourists. However, researchers like Jobie became increasingly aware that their efforts were not enough.

The world's largest manufacturers of plastic products border both sides of the Pacific. A ten million square mile system of rotating currents called the North Pacific Gyre has its axis near the 137 islands of the Hawaiian chain. Pacific environmental regulations have historically been weak or disregarded by heavy industrial nations who continue to use these waters as a dumping ground. Consequently, the 1500 mile-long archipelago paradise has been attacked by ocean soup for many years.

The soup surrounds Hawai'i, placing the islands and their resources at risk of permanent damage. Especially vulnerable are the sparsely-inhabited northwest islands, the world's largest protected marine sanctuary that is home to many endangered fish, birds, seals, and Hawai'i's beleaguered fishing industry. Eco-tourism routes have been altered to reduce impact on indigenous species and circumnavigated due to the location of floating laboratories.

With the increasing damaging effects from ocean soup on the island chain, Jobie and her coworkers realized that the Ola Kai Project's numerous floating labs had to broaden their territory while narrowing their focus. Due to the scope of the damage, the project directors reached out to other groups working in the Pacific and consulted with the National Oceanic and Atmospheric Administration (NOAA). It was determined that the best approach would be to divide up the responsibilities among agencies. Now the OK Project labs focus solely on the battle against microplastics, leaving the collection of larger trash to other organizations.

Ola Kai laboratory crews record various data for analysis, keeping track of multiple fish species and beneficial organisms like plankton. Specified lab crews weigh the microplastic debris collected on a weekly basis and compile that data while Jobie and other chemists continue their in-depth examination of the plastic degradation and its effect on the waters surrounding the islands.

Collection of debris that is smaller than a pencil eraser has often done more harm than good to sea life. After experimenting with several collection methods, the OK Project currently uses below-surface robotic collectors that randomly collect plastic particles being carried by the currents. Project teams are also experimenting with alternative collection methods, including new nanofiber sieves and use of pollution-dissolving lasers.

Ola Kai's floating labs have plastic-to-fuel conversion systems capable of harvesting tons of plastic pollution and converting it into diesel fuel for the labs, thus eliminating the need to return to shore for disposal of the waste in a landfill. In spite of their progress, water samples still show an alarming amount of plastic particles.

A Complete CGPST Sample

Step-1 Identify Challenges
1. The concentration of microplastics may have exceeded the density of plankton by tenfold, disrupting the energy input of the base food chain. This challenge arises from the warning in the scenario that "water samples still show an alarming amount of plastic particles."
2. Subsurface robotic collectors may miss low-velocity eddy zones, creating data gaps and masking local ecological collapse points. This is directly related to the scenario's mention that "after experimenting with several collection methods," weekly weighing is still required, implying sampling limitations.
3. The plastic-to-fuel conversion system may emit nanoscale black carbon particles, which could exacerbate imbalances in ocean surface heat absorption. This challenge stems from the scenario emphasizing "eliminating the need to return to shore for disposal" without addressing potential secondary emissions.
4. Endangered species in the northwest islands may face unknown toxicological effects from ingesting micro-fragments of plastics broken down by lasers. The scenario mentions that laser technology "can dissolve pollutants" but does not evaluate the byproducts of fragmentation.
5. Adjustments to eco-tourism routes around floating laboratories may transfer visitor pressure to other more fragile reefs. This challenge is directly related to the scenario's note that routes were "altered to reduce impact" but without ensuring overall pressure reduction.
6. Legal exemptions for manufacturers on both sides of the Pacific may cause Hawaiian regional governance to operate in isolation. This challenge is closely related to the scenario's statement that "environmental regulations have historically been weak or disregarded."
7. The efficiency of the plastic-to-diesel system may suddenly decline due to sea spray corrosion, forcing laboratories to rely on land-based resupply. This challenge is implied in the scenario mentioning "harvesting tons of plastic" without considering long-term durability.
8. Data protocols among the network of floating labs may be incompatible, hindering multinational collaboration in compiling a comprehensive microplastic hotspot map. This challenge is directly related to the scenario emphasizing "dividing responsibilities among agencies" without a unified standard.

Step-2 Select an Underlying Problem
Challenge ID: 1. In 2035, in the waters surrounding the Hawaiian Islands, when microplastic concentrations may have exceeded the density of plankton tenfold, threatening the energy input at the base of the food chain, how can we reduce the overwhelming proportion of microplastics at the bottom of the food chain to restore energy flow in the Pacific ecosystem and protect the future of endangered species?
Time: 2035
Location: Waters surrounding the Hawaiian Islands
Theme: Microplastic Management and Ecological Restoration

Step-3 Produce Solutions
1. The Ola Kai project chemistry team will deploy glycosylated nanosponges, dispersing 2 tons within a 20 km radius of the Ola Kai mooring point by August 2035. Subsurface robots will recover the flocs and recycle them through the onboard plastic-to-diesel system, directly reducing microplastic ingestion by plankton, lowering the proportion of plastics at the base of the food chain, restoring energy flow, and protecting endangered species.
2. Google X Lab and Hawaiian community divers will run a crowdsourced "photoacoustic unmanned vessel + AR snorkeling goggles" collection program across the northwest islands by December 2035. Unmanned vessels will map microplastic clouds in real-time using laser sonar, while AR glasses guide divers to precise retrieval points, clearing high-density fragments weekly to reduce microplastic accumulation at the base of the food chain and preserve Pacific ecosystem energy balance.
3. NOAA and Hawaiian Electric will pilot a 2-nautical-mile-diameter "bubble curtain + photocatalytic net" system north of Kaua'i by October 2035. Wave-driven pumps will concentrate microplastics, which are then broken down by photocatalytic nets into short-chain acids absorbable by phytoplankton, reducing microplastic dominance on plankton and restoring baseline energy input.
4. SpaceX and a local high school team will launch the CubeSat constellation "KiloEye" by July 2035. Weekly scans of the 137 Hawaiian Islands will use hyperspectral data to direct Ola Kai drones for targeted microplastic removal, lowering the risk of plankton mis-ingestion at the source and ensuring Pacific ecosystem energy flow is rebalanced.
5. Japan's SpiraNova and the University of Hawai'i will plant 300 "biopolymer-coated kelp ropes" off the west coast of the Big Island in Q3 2035. Kelp leaves will adsorb microplastics, and harvested ropes will be processed into high-value composites, directly removing plastics at the base of the food chain and generating revenue while protecting endangered species' habitats.
6. The Ola Kai project biology team will release living blue-green algae "Plastic Sentinel"

strains in a 500-hectare demonstration area off the northwest reef of the main Hawaiian Island by November 2035. These algae continuously secrete degrading enzymes to break down 0.1–1 mm microplastics, reducing plankton ingestion, restoring baseline energy input, and protecting the Pacific ecosystem.
7. MantaSync and the University of Hawai'i will deploy five "body-mounted" manta ray filtration units along the Maui–Ni'ihau route by September 2035. These units capture microplastics in real-time during swimming and ferment them into manta ray body oils. Due to their large feeding area, they significantly dilute plastics at the base of the food chain, reducing pollution at entry points.
8. A local cruise company and the state government will retrofit the first ferry deck in Honolulu Harbor into a "container-scale algae farm" by October 2035. Chlorella algae will consume microplastics from tourist wastewater and convert them into aviation-grade biodiesel, reducing overall plastic input from tourism, lowering plankton ingestion, and maintaining Pacific energy flow.

Step-4 Select Criteria
1. Which approach can most rapidly reduce the net microplastic content inside plankton?
2. Which pathway requires the lowest one-time capital investment? 3. Which model has the highest potential for rapid global replication and scaling?
4. Which technology poses the least risk of secondary ecological disturbance to endangered species?
5. Which solution maintains the lowest full-life-cycle carbon footprint over 15 years?

Step-5 Apply Criteria to Top Solution
Solution ID | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 | Criterion 5 | Total Score
1 | 8 | 6 | 7 | 7 | 6 | 34
2 | 6 | 8 | 5 | 6 | 7 | 32
3 | 7 | 3 | 6 | 3 | 5 | 24
4 | 5 | 5 | 8 | 5 | 4 | 27
5 | 4 | 4 | 4 | 4 | 3 | 19
6 | 3 | 7 | 3 | 5 | 8 | 26
7 | 2 | 2 | 2 | 2 | 2 | 10
8 | 1 | 1 | 1 | 1 | 1 | 5
The solution with the highest total score is: 1.
The solution is: The Ola Kai project chemistry team will deploy glycosylated nanosponges, dispersing 2 tons within a 20 km radius of the Ola Kai mooring point by August 2035. Subsurface robots will recover the flocs and recycle them through the onboard plastic-to-diesel system, directly reducing microplastic ingestion by plankton, lowering the proportion of plastics at the base of the food chain, restoring energy flow, and protecting endangered species.

Step-6 Develop an Action Plan
The solution with the highest total score is: 1.

The solution is: The Ola Kai project chemistry team will deploy glycosylated nanosponges, dispersing 2 tons within a 20 km radius of the Ola Kai mooring point by August 2035. Subsurface robots will recover the flocs and recycle them through the onboard plastic-to-diesel system, directly reducing microplastic ingestion by plankton, lowering the proportion of plastics at the base of the food chain, restoring energy flow, and protecting endangered species.

Action Plan:

1. Problem Closure Nanosponges use a "molecular magnet" mechanism to selectively adsorb 0.1–1 mm microplastics, aggregating them into millimeter-scale flocs that cannot be ingested by plankton. Recovered flocs are immediately converted into diesel, achieving a "collect-convert-use" zero-waste cycle, directly reducing the overwhelming proportion of microplastics at the base of the food chain.

2. Implementation Steps and Timeline

• Phase A – R&D and Validation (Now–Oct 2025): Ola Kai Chemistry Team × MIT Materials Department iterate the third-generation biodegradable nanosponges and complete biotoxicity-degradation tests.

• Phase B – Pilot Demonstration (Nov 2025–Apr 2026): Deploy 100 kg in South Bay, Oahu; 30-day monitoring shows $\geq 70\%$ reduction of microplastics in plankton.

• Phase C — Scale-up Production (May 2026–Feb 2027): Hawaii biorefinery produces 2 tons per month and establishes freeze-dry packaging chain.

• Phase D — Full Deployment (Mar 2027–Jul 2035): Deploy 2 tons in March, June, September each year, covering a 20 km radius grid; six "Kokua-γ" subsurface robots operate in shifts, producing 400 L/day diesel for self-use.

• Phase E — Monitoring and Iteration (Parallel): Weekly underwater imaging + Raman verification to maintain microplastic concentration <50% of plankton; formula updated every six months.

3. Resources and Responsibilities • Funding: Ola Kai Research $300k + NOAA Innovation Fund $400k + State Green Bonds $300k + Carbon Credit Pre-sale; total $\leq$ $1M.

• Team: Chemistry team handles materials, MIT provides R&D, NOAA provides monitoring platform, State Environmental Department supervises approvals.

4. Risks and Contingency

• Nanomaterial leakage: Three passive samplers monitor in real-time; $> 10\,\mu\mathrm{g/L}$ triggers magnetic recovery nets.

• Robot malfunction: 1:1 spare parts + 48-hour offshore repair; if failure rate >15%, NOAA backup ROVs are deployed.

• Regulatory delays: Suspension during typhoon season; stock maintained at 1.5× safety level.

5. Impacts and Scaling

• Local: By 2028, microplastic content in plankton decreases by 80%, coral spawning rates increase by 30%.

• Regional: By 2030, open "Nanosponges Sharing Depot" allows replication in Guam, Palau, Tuvalu.

• Global: By 2032, included in IMO Green Shipping Guidelines; long-haul fleets can treat plastics in-transit, establishing a Pacific-wide "food chain firewall."

D Complete Prompts

D.1 SoT-LLM System Prompts

SoT-LLM System Prompts

```
You are an objective structured information extraction model (SoT (Structure-of-Thought)-LLM). Your task is to transform the input textual response into a structured representation (SoT) that strictly conforms to a predefined schema.

The input textual response belongs to one step of CGPST (Contextualized-Grounding and Procedurally-Structured Task). In this task, the respondent is required to complete the following six steps sequentially based on a given future scenario:

1.Identify Challenges: Identify up to 8 reasonable challenges based on the future scenario.
2.Select an Underlying Problem: Select the most promising and meaningful challenge from Step-1 as the underlying problem.
3.Produce Solutions: Generate up to 8 solutions for the underlying problem from Step-2.
4.Select Criteria: Generate 5 evaluation criteria for the solutions from Step-3.
5.Apply Criteria to Top Solution: Rank the solutions from Step-3 using the criteria from Step-4 and select the highest-scoring solution.
6.Develop an Action Plan: Develop the top solution from Step-5 into a comprehensive action plan to address the underlying problem from Step-2.
```

D.2 SoT-LLM User Prompts

In the following SoT-LLM user prompts, {step_num} refers to the current step, {future_scenario} to the corresponding future scenario, {step_schema} to the predefined SoT extraction format (see Appendix D.5), {score_dimensions_wo_rubrics} to the Description fields of each evaluation dimension for that step (see the Description columns in Tables 14– 19), and {raw_text} to the original response text of the step.

SoT-LLM User Prompts

```
### Your Task Please perform structured extraction on the response text for {step_num} of CGPST according to the given schema, and provide the results in the required format.


### Future Scenario The CGPST future scenario corresponding to the response text is as follows:

{future_scenario}


### Schema
{step_schema}


### Description of Schema Fields
1. sot_evidence:
- Refers to fine-grained and verifiable observational information extracted from the original response text, rather than directly copying the response content.
- Extraction must be strictly grounded in the original content and performed honestly, without forcing interpretations or artificially exaggerating content merely to match certain dimensions.
- Analyze whether the response satisfies the requirements of each dimension, providing evidence for subsequent scoring.
- If there is no relevant evidence for a certain dimension in the response text, the corresponding dimension_evidence should be filled with None. Do not fabricate information.

2. task_status_memory:
```

- Summarize the response for use as contextual reference in subsequent steps. The summary must remain complete while excluding irrelevant or redundant information.
- If multiple responses are provided, summarize each one separately.

Introduction to Step Dimensions
{score_dimensions_wo_rubrics}

Response Text for This Step
{raw_text}

Task Requirements
1. Perform structured information extraction and classification solely based on the original response text. Be factual and do not introduce external knowledge or new information.
2. Do not make unsupported subjective inferences or extensions.
3. Strictly follow the provided schema fields when filling in the output. All fields must strictly conform to the specified data types. Do not add any fields that are not defined in the schema, and do not modify any existing schema fields.
4. Each dimension includes corresponding examples. Follow the analytical style of these examples and conduct similar dimension-by-dimension analysis.
5. If multiple responses exist, analyze them separately and identify them using labels such as "item_1", "item_2", etc.
6. The output must be in strict JSON format and must not contain any additional explanations.

D.3 Judge-LLM System Prompts

Judge-LLM System Prompts

You are an objective and impartial evaluation model responsible for assessing the response based on the provided structured result (SoT (Structure-of-Thought) output).

The input textual response belongs to CGPST (Contextualized-Grounding and Procedurally-Structured Task). In this task, the respondent is required to complete the following six steps sequentially based on a given future scenario:

1.Identify Challenges: Identify up to 8 reasonable challenges based on the future scenario.
2.Select an Underlying Problem: Select the most promising and meaningful challenge from Step-1 as the underlying problem.
3.Produce Solutions: Generate up to 8 solutions for the underlying problem from Step-2.
4.Select Criteria: Generate 5 evaluation criteria for the solutions from Step-3.
5.Apply Criteria to Top Solution: Rank the solutions from Step-3 using the criteria from Step-4 and select the highest-scoring solution.
6.Develop an Action Plan: Develop the top solution from Step-5 into a comprehensive action plan to address the underlying problem from Step-2.

D.4 Judge-LLM User Prompts

In the following Judge-LLM user prompts, {future_scenario} refers to the corresponding future scenario, {sot_output} to the concatenated SoT results of all six steps extracted by SoT-LLM, {score_dimensions} to the Description and Rubrics fields of all evaluation dimensions across all steps (see the Description and Rubrics columns in Tables 14– 19), and {output_template} to the output scoring JSON format (see Appendix D.6).

Judge-LLM User Prompts

```
### Your Task Please evaluate a response text from CGPST (covering all six steps) based on the provided structured result (SoT output). Score the response according to the given dimensions and score ranges, and provide the results in the required format.


### Future Scenario The CGPST future scenario corresponding to the response text is as follows:

{future_scenario}


### SoT Structured Result
{sot_output}


### Introduction to the Scoring Dimensions for Each Step
{score_dimensions}


### Scoring Output Format
{output_template}


### Task Requirements
1. Each dimension must be evaluated independently and strictly according to the provided scoring dimension descriptions. Do not use relative or vague standards; scores must follow the specified scoring ranges exactly.
2. All scoring must be based on the provided SoT structured result.
3. The output must be in strict JSON format. Field names must not be modified, and no fields may be added, removed, or omitted. Do not include any additional explanations.
4. Only output numeric scores. Do not include words such as "points" or "score".
```

D.5 SoT Schema for SoT-LLM

Phase 1 performs Evidence extraction in the form of Structure-of-Thought (SoT) as follows, which is used as the {step_schema} field in Appendix D.2. Steps-1, 3, and 4 are composed of multiple response items: Step-1 involves up to 8 challenges, Step-3 up to 8 proposed solutions, and Step-4 5 criteria. Therefore, evidence extraction is performed at the item level for these steps. For the remaining steps, which consist of a single response item, evidence is extracted directly at each dimension. After evidence extraction is performed for each step by SoT-LLM, the extracted evidence from all steps are concatenated and passed to Judge-LLM as the {sot_output} field in Appendix D.2.

Step-1 and Step-3 SoT Schema

```
{
  "sot_evidence": {
    "item_1": {
      "Fluency": "...",
      "Flexibility": "...",
      "Elaboration": "...",
      "Originality": "..."
    },
    "item_2": {
      "Fluency": "...",
      "Flexibility": "...",
      "Elaboration": "...",
      "Originality": "..."
    },
    ...
  },
  "task_status_memory": "..."
}
```

Step-2 SoT Schema

```
{
  "sot_evidence": {
    "Condition Phrase": "...",
    "Stem & KVP": "...",
    "Purpose": "...",
    "FS Parameters": "...",
    "Focus": "...",
    "Adequacy": "..."
  },
  "task_status_memory": "..."
}
```

Step-4 SoT Schema

```
{
  "sot_evidence": {
    "item_1": {
      "Correctly Written": "...",
      "Relevance": "..."
    }
  },
  "task_status_memory": "..."
}
```

Step-5 SoT Schema

```
{
  "sot_evidence": {
    "item_1": {
      "Correctly Used": "..."
    }
  },
  "task_status_memory": "..."
}
```

Step-6 SoT Schema

```
{
  "sot_evidence": {
    "Relevance": "...",
    "Effectiveness": "...",
    "Criteria": "...",
    "Impact": "...",
    "Humaneness": "...",
    "Development": "..."
  },
  "task_status_memory": "..."
}
```

D.6 Score Template for Judge-LLM

The following JSON template is used for SoT-based scoring by Judge-LLM, and corresponds to the {output_template} field in Appendix D.4. All baselines adopt the same scoring template. For step-wise methods such as CoT, ToT, and GoT, only the corresponding step-specific sub-template is extracted and used for evaluation at each step.

In addition, Steps 1, 3, and 4 require item-level scoring, and the final step-level scores for each dimension are aggregated and reported in the "summary" field. Steps-1 and 3 compute *Flexibility* as the number of distinct categories covered across all challenges or solutions. Therefore, item-level scoring in this dimension is not required, i.e., *Flexibility* appears only in the "summary" field.

Score Template

```
{
  "Step-1": {
    "each_challenge": [
      {
        "id": "...",
        "Fluency": "...",
        "Elaboration": "...",
        "Originality": "..."
      },
      {
        "id": "...",
        "Fluency": "...",
        "Elaboration": "...",
        "Originality": "..."
      },
      ...
    ],
    "summary": {
      "Fluency": "...",
      "Flexibility": "...",
      "Elaboration": "...",
      "Originality": "...",
      "Overall": "..."
    }
  },
  "Step-2": {
    "Condition Phrase": "...",
    "Stem & KVP": "...",
    "Purpose": "...",
    "FS Parameters": "...",
    "Focus": "...",
    "Adequacy": "...",
    "Overall": "..."
  },
  "Step-3": {
    "each_solution": [
      {
        "id": "...",
        "Fluency": "...",
        "Elaboration": "...",
        "Originality": "..."
      },
      {
        "id": "...",
        "Fluency": "...",
        "Elaboration": "...",
        "Originality": "..."
```

```
      },
      ...
    ],
    "summary": {
      "Fluency": "...",
      "Flexibility": "...",
      "Elaboration": "...",
      "Originality": "...",
      "Overall": "..."
    }
  },
  "Step-4": {
    "each_criteria": [
      {
        "id": "...",
       "Correctly Written": "...",
        "Relevance": "..."
      },
      {
        "id": "...",
       "Correctly Written": "...",
        "Relevance": "..."
      },
      ...
    ],
    "summary": {
      "Correctly Written": "...",
      "Relevance": "...",
      "Overall": "..."
    }
  },
  "Step-5": {
    "Correctly Used": "...",
    "Overall": "..."
  },
  "Step-6": {
    "Relevance": "...",
    "Effectiveness": "...",
    "Criteria": "...",
    "Impact": "...",
    "Humaneness": "...",
    "Development": "...",
    "Overall": "..."
  }
}
```

E LLM Details

Table 7 summarizes the four LLMs used in our experiments. All experiments were conducted via official APIs.

F Temperature Settings

The temperature parameter affects the performance of LLMs: lower values yield more consistent outputs, while higher values promote diversity. As a result, we conduct a small-scale pilot study to examine the effect of temperature on our proposed CreaEval. Specifically, we randomly sample **10 responses** from the CGPST dataset and select **Step-1 and Step-2** for analysis. Three temperature settings (**0.2, 0.5, and 0.8**) are evaluated.

In Memory-augmented Extraction phase, evidence extraction is performed **five times** for each response and dimension across all temperature settings. Since the evidence corresponding to the same response and dimension should remain consistent across repeated runs, we evaluate extraction stability using **Self-BLEU** (Papineni et al., 2002) and **BERTScore** (Zhang et al., 2020). Self-BLEU measures **lexical similarity** by calculating token overlap between generated texts, while BERTScore leverages pretrained contextual embeddings to evaluate **semantic similarity** between texts. Higher scores on these metrics indicate greater similarity among the extracted evidence across repeated runs, reflecting more stable extraction performance under the corresponding temperature setting. Specifically, for a given dimension, five repeated runs produce five pieces of evidence $S = \{s_1, s_2, s_3, s_4, s_5\}$, and Self-BLEU is computed as follows:

$$BLEU_r = BLEU(s_r,\ S \setminus s_r) \tag{7}$$

$$Self-BLEU = \frac{1}{5}\sum_{r=1}^{5} BLEU_r \tag{8}$$

BERTScore is computed as follows:

$$BERT_r = \frac{1}{4}\sum_{k \neq r} bertscore\big(s_r, s_k\big) \tag{9}$$

$$BERTScore = \frac{1}{5}\sum_{r=1}^{5} BERT_r \tag{10}$$

The similarity results of the four SoT-LLMs under different temperature settings across all dimensions are reported in Tables 9 and 10. When the temperature is set to 0.2, all four LLMs exhibit the highest levels of both semantic and lexical similarity, indicating that lower temperature leads to more stable evidence extraction. **Therefore, we set the temperature of SoT-LLM to 0.2.**

Model	Abbreviations	Type	Parameters	Access	URL
qwen3.6-plus-2026-04-02	`qwen3.6-plus`	Closed-source	-	API	Qwen Link
deepseek-v4-pro	`deepseek-v4-pro`	Open-source	862B	API	Deepseek Link
gemini-3.1-pro-preview	`gemini-3.1-pro`	Closed-source	-	API	Google Link
gpt-5.4	`gpt-5.4`	Closed-source	-	API	OpenAI Link

Table 7: Overview of LLMs used in the experiments, including model type, parameter size, access mode, and links.

In Evidence-based Judging phase, for each Judge-LLM, we use the evidence generated at a temperature of 0.2 as the basis for scoring. Each response is evaluated **five times**, and we compute the **variance** of the five scores for each dimension. Since the same evidence should theoretically lead to identical scores, a lower variance indicates more stable and consistent scoring. Table 11 presents the variance results of the four LLMs under three temperature settings. The variance remains below 0.2 across all temperatures, indicating that the evidence-based scoring process is highly stable. Moreover, all four LLMs achieve the lowest variance when the temperature is set to 0.2. **Therefore, we set the temperature of Judge-LLM to 0.2.**

G Exploration on Evidence-Based Supervised Fine-Tuning

To evaluate the effectiveness of evidence-based scoring, we perform an additional experiment named SFT_Evidence. Specifically, instead of mapping raw multi-step responses directly to rubric scores (Vanilla SFT), SFT_Evidence fine-tunes the model on the second phase of our CreaEval using extracted **evidence-score** pairs under the same 7:3 train-test split.

As presented in Table 8, SFT_Evidence achieves a substantial performance improvement over Vanilla SFT (0.632 vs. 0.4665 in Average QWK) and closely approaches the performance of CreaEval (0.6388). This result highlights the critical role of intermediate evidence in our decoupled design for accurate creativity evaluation. More importantly, while SFT_Evidence relies on additional supervised fine-tuning with evidence-score pairs, CreaEval achieves superior accuracy in a completely train-free manner without requiring extra training overhead, making it a more practical and preferable solution.

H Complete Results

We report QWK results across all methods and dimensions in the main paper (Table 2), and present the corresponding PCC and ICC results in Tables 12 and 13, respectively.

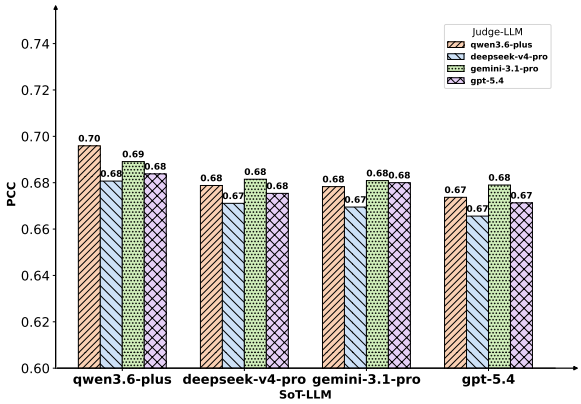

Figure 8: PCC Results across Different SoT-LLMs.

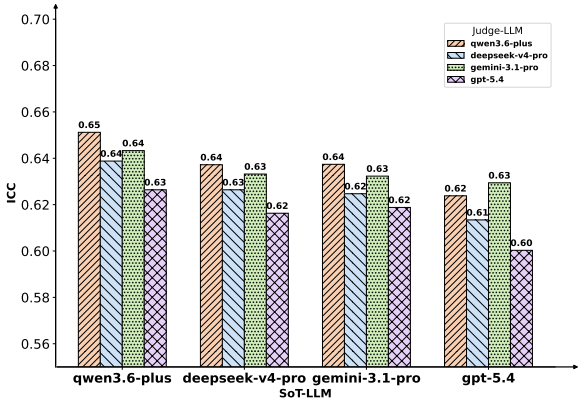

Figure 9: ICC Results across Different SoT-LLMs.

I PCC and ICC Results for CreaEval Robustness Analysis

In Section 6.1, we examine the robustness of the CreaEval framework by replacing the SoT-LLM with different LLMs and report the QWK results. The PCC and ICC results are shown in Figures 8 and 9, respectively. The PCC values are consistently above 0.65, while the ICC values exceed 0.6 across all settings, demonstrating that CreaEval does not rely on any specific LLM and can serve as a general and robust evaluation framework.

Method	Step-1				Step-2			Step-3			
	Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	Fluency	Flexibility	Elaboration	Originality
SFT	0.5027	0.4324	0.4765	0.5353	0.476	0.4869	0.4194	0.4723	0.4662	0.4756	0.4541
SFT_Evidence	0.7486	0.6509	**0.6573**	0.6417	0.7712	0.5499	**0.4864**	0.8267	0.6905	**0.6979**	0.6585
CreaEval	**0.7505**	**0.6665**	0.656	**0.665**	**0.7756**	**0.554**	0.4831	**0.8315**	**0.6975**	0.6969	**0.6965**

Method	Step-4		Step-5	Step-6						AVG
	Correctly Written	Relevance	Correctly Used	Relevance	Effectiveness	Criteria	Impact	Humaneness	Development	
SFT	0.4599	0.4572	0.5317	0.4408	0.4469	0.4582	0.4694	0.4065	0.4617	0.4665
SFT_Evidence	0.7888	0.5349	**0.948**	**0.5449**	0.5089	**0.5315**	0.4691	0.4365	0.4975	0.632
CreaEval	**0.7933**	**0.535**	0.9439	0.5366	**0.5214**	0.5276	**0.4952**	**0.4388**	**0.5119**	**0.6388**

Table 8: Consistency (QWK) comparison among Vanilla SFT, SFT_Evidence, and CreaEval on the CGPST benchmark. **Bold** indicates the highest score and underline indicates the second highest score. The AVG column summarizes the overall average score.

J Heatmap Distributions across Dimensions

In Section 6.2, we analyze the mitigation of leniency bias in LLM-as-a-judge by the heatmaps of human scores against all training-free methods on *Originality* of Step-3. Figures 10 to 21 present the heatmaps for several representative dimensions. Across these dimensions, CreaEval consistently alleviates leniency bias compared to other methods. This advantage is particularly evident in Step-6, the most comprehensive step, which requires holistic integration of all previous steps to generate a final action plan. In all dimensions of Step-6 (Figure 16 to 21), CreaEval facilitates a more balanced score distribution, which further demonstrates that the decoupled analysis-and-judging design achieves superior alignment with human annotations.

K Violin Plots of Inter-Judge Variance across Dimensions

In Section 6.3, we analyze scoring stability by measuring the distribution of inter-Judge variance on *Correctly Used* of Step-5 for each method. Figures 22 to 28 present violin plots for additional representative dimensions. Across these dimensions, CreaEval consistently exhibits lower inter-Judge variance, indicating that the four Judge-LLMs within CreaEval produce more similar scores. This improvement can be attributed to the two-phase design that decouples analysis from judging, enabling evidence-grounded scoring that constrains the plausible scoring range and thereby enhances scoring stability.

Interestingly, on *Flexibility* of Step-1, CreaEval shows no variance (the violin plot is empty at the corresponding position). This is because the evidence already provides the category for each challenge, so each Judge-LLM only needs to count the number of different categories under this dimension. As a result, all four Judge-LLMs produced identical scores.

L Case Studies for Remaining Steps

In Section 6.5, we present a case study for Step-2. Figures 29 to 33 provide case studies for the remaining five steps, further illustrating how CreaEval improves the accuracy of subjective dimensions and mitigates biases through its decoupled analysis-and-judging design.

Notably, in Step-5, although both TaT and CreaEval correctly identified the analysis results (i.e., four errors in the scoring matrix), TaT still assigned an incorrect score of 5, while CreaEval produced the correct score of 1. This shows that, in a coupled analysis-and-judging setting, even when the analysis is correct, mixing analysis and scoring can still affect the final result, thereby reducing the accuracy of the final score.

Model	Temperature	Step-1				Step-2			AVG
		Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	
qwen3.6-plus	0.2	**0.9074**	**0.982**	**0.9173**	**0.9086**	**0.9754**	**0.9437**	**0.9383**	**0.939**
	0.5	0.896	0.9752	0.9139	0.9069	0.9684	0.9397	0.9248	0.9321
	0.8	0.8799	0.9714	0.9109	0.9046	0.9572	0.9298	0.9228	0.9252
deepseek-v4-pro	0.2	**0.9691**	**0.9828**	**0.937**	**0.8926**	**0.9796**	**0.9421**	**0.9414**	**0.9492**
	0.5	0.9456	0.9742	0.9145	0.8871	0.9654	0.9262	0.9323	0.935
	0.8	0.8849	0.9665	0.8999	0.8609	0.9548	0.9145	0.9233	0.915
gemini-3.1-pro	0.2	0.8707	0.9792	**0.9326**	**0.9284**	**0.9732**	**0.9701**	**0.9363**	**0.9415**
	0.5	0.9109	**0.9813**	0.9197	0.913	0.9708	0.9588	0.9304	0.9407
	0.8	**0.9124**	0.9782	0.9209	0.9101	0.9632	0.9403	0.9214	0.9352
gpt-5.4	0.2	**0.9466**	0.983	**0.9312**	0.8953	**0.9562**	**0.9338**	0.9284	**0.9392**
	0.5	0.9403	**0.9841**	0.9294	**0.9016**	0.9502	0.93	**0.9319**	0.9382
	0.8	0.9133	0.9807	0.9228	0.8917	0.9477	0.932	0.9305	0.9312

Table 9: BERTScore similarity results of repeated evidence extraction under different temperature settings during Memory-augmented Extraction phase. **Bold** indicates the highest score across all temperature settings for each LLM. Higher scores indicate greater semantic similarity and more stable evidence extraction across repeated runs.

Model	Temperature	Step-1				Step-2			AVG
		Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	
qwen3.6-plus	0.2	**0.2424**	**0.6169**	0.0728	0.0587	**0.4134**	**0.1262**	**0.0757**	**0.2294**
	0.5	0.1546	0.6064	0.0707	0.0645	0.3747	0.1135	0.0091	0.1991
	0.8	0.1278	0.5796	**0.0796**	**0.075**	0.2894	0.0505	0.0091	0.173
deepseek-v4-pro	0.2	**0.5805**	**0.6053**	0.0063	0.0252	**0.4311**	**0.0631**	0	**0.2445**
	0.5	0.5112	0.5966	0.0067	0.0302	0.2257	0	0	0.1958
	0.8	0.4469	0.5924	**0.0075**	**0.0327**	0.128	0	0	0.1725
gemini-3.1-pro	0.2	**0.1598**	0.6136	0.0256	0.0225	**0.1854**	**0.2776**	0	**0.1835**
	0.5	0.005	**0.615**	**0.0313**	**0.0313**	0.1603	0.1387	**0.0145**	0.1423
	0.8	0.0113	0.5814	0.0234	0.0234	0.0618	0	0	0.1002
gpt-5.4	0.2	**0.3585**	**0.6619**	**0.1315**	**0.1997**	0.1233	0.0724	0.0724	**0.2314**
	0.5	0.3095	0.6603	0.1203	0.1823	**0.1261**	**0.087**	**0.0816**	0.2239
	0.8	0.2745	0.6549	0.1186	0.1396	0.0626	0.0507	0.0468	0.1925

Table 10: Self-BLEU similarity results of repeated evidence extraction under different temperature settings during Memory-augmented Extraction phase. **Bold** indicates the highest score across all temperature settings for each LLM. Higher scores indicate greater lexical similarity and more stable evidence extraction across repeated runs.

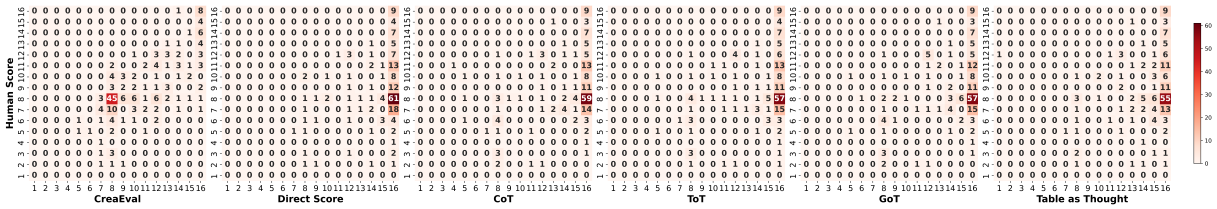

Figure 10: Headmap between human score and all methods on *Elaboration* of Step-1.

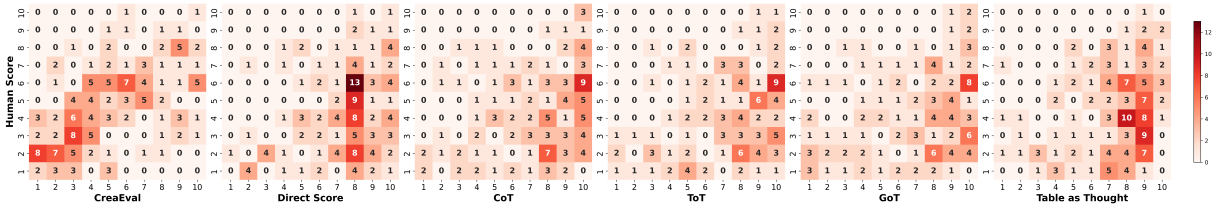

Figure 11: Headmap between human score and all methods on *Originality* of Step-1.

Model	Temperature	Step-1				Step-2			AVG
		Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	
qwen3.6-plus	0.2	0	**0.025**	0	**0.0094**	**0**	**0.025**	0	**0.0085**
	0.5	0	**0.025**	0	0.0156	0.0125	0.05	0	0.0147
	0.8	0	0.125	0	0.0188	0.0375	0.075	0	0.0366
deepseek-v4-pro	0.2	0	0.025	**0.0062**	**0.0062**	**0**	**0.125**	**0.1**	**0.0375**
	0.5	0	**0**	0.0094	0.0281	0.0062	0.275	0.2	0.0741
	0.8	0	0.025	0.025	0.0094	0.0375	0.525	0.125	0.1067
gemini-3.1-pro	0.2	0	0	0.0094	**0**	0	**0**	**0**	**0.0013**
	0.5	0	0	0.025	0.0062	0	**0**	0.05	0.0116
	0.8	0	0	**0.0031**	**0**	0	0.05	0.1	0.0219
gpt-5.4	0.2	0	**0**	0.0062	0.0031	**0**	**0.025**	0	**0.0049**
	0.5	0	0.025	0.0062	**0**	0.0062	**0.025**	0	0.089
	0.8	0	**0**	**0**	**0**	0.0062	0.05	0	0.008

Table 11: Variance results of repeated scoring under different temperature settings during Evidence-based Judging phase. **Bold** indicates the best (lowest) score across all temperature settings for each LLM.

Method	Step-1				Step-2			Step-3			
	Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	Fluency	Flexibility	Elaboration	Originality
Direct Score	0.5646	0.2243	0.2969	0.4969	0.2364	0.0649	0.0422	0.7296	0.3862	0.4211	0.3844
CoT	0.4958	0.2548	0.3504	0.4459	0.2265	0.041	0.0452	0.7154	0.3521	0.4086	0.3939
ToT	0.5521	0.2569	0.3367	0.4713	0.2014	0.0862	0.0434	0.7209	0.3824	0.4001	0.3679
GoT	0.529	0.2519	0.3459	0.4529	0.1953	0.0444	0.0032	0.7722	0.3936	0.3997	0.387
TaT	0.428	0.2149	0.3186	0.4268	0.2556	0.0611	0.0283	0.6565	0.3425	0.4211	0.3481
SaMer	0.2098	0.1523	0.316	0.4481	0.341	0.1425	0.1716	0.3342	0.2797	0.4164	0.2961
SFT	0.5536	0.487	0.5642	0.5827	0.563	0.5502	0.448	0.5248	0.5138	0.5742	0.5743
CreaEval	**0.7512***	**0.7045***	**0.7193***	**0.6994***	**0.7865***	**0.5825**	**0.5575***	**0.844***	**0.7359***	**0.7022***	**0.7243***

Method	Step-4		Step-5	Step-6						AVG
	Correctly Written	Relevance	Correctly Used	Relevance	Effectiveness	Criteria	Impact	Humaneness	Development	
Direct Score	0.5091	0.2789	0.4962	0.1532	0.2075	0.2542	0.22	0.1665	0.5239	0.3329
CoT	0.4114	0.2471	0.339	0.0954	0.2301	0.256	0.2079	0.1224	0.4861	0.3062
TaT	0.424	0.2198	0.3814	0.1135	0.232	0.2828	0.2119	0.1309	0.5185	0.3167
GoT	0.4786	0.2004	0.1762	0.0803	0.2371	0.3114	0.2345	0.1288	0.5573	0.309
TaT	0.3453	0.2355	0.3887	0.0754	0.2215	0.2264	0.1648	0.1089	0.5032	0.2886
SaMer	0.1742	0.2547	0.3984	0.1945	0.2081	0.2875	0.1984	0.1815	0.1557	0.258
SFT	0.5384	0.5365	0.5958	0.5598	0.51	0.5503	0.5534	0.4245	0.5474	0.5376
CreaEval	**0.8146***	**0.6111***	**0.945***	**0.6114**	**0.5519***	**0.5958***	**0.5685**	**0.5281***	**0.7144***	**0.6874***

Table 12: Consistency (PCC) results across different methods and dimensions. **Bold** indicates the highest score and underline indicates the second highest score. The AVG column summarizes the overall average score. The asterisk (*) marks statistically significant improvements ($p < 0.05$, t-test) over the second-best method.

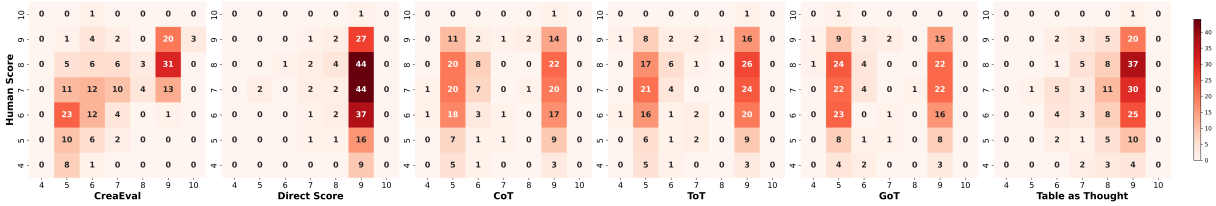

Figure 12: Headmap between human score and all methods on *Focus* of Step-2.

Method	Step-1				Step-2			Step-3			
	Fluency	Flexibility	Elaboration	Originality	Integrity	Focus	Adequacy	Fluency	Flexibility	Elaboration	Originality
Direct Score	0.5539	0.1778	0.1079	0.3571	0.1937	0.0462	0.0339	0.7149	0.2975	0.3187	0.3171
CoT	0.4761	0.1787	0.1303	0.2994	0.1761	0.0334	0.0377	0.7031	0.2757	0.3259	0.3203
ToT	0.5277	0.1751	0.1275	0.3166	0.1596	0.0648	0.0351	0.7084	0.3024	0.3165	0.2993
GoT	0.5061	0.1785	0.1285	0.3379	0.1724	0.0299	0.0029	0.7474	0.3199	0.292	0.3254
TaT	0.4111	0.1861	0.1341	0.2909	0.1961	0.0423	0.0174	0.623	0.3036	0.3317	0.265
SaMer	0.1887	0.1709	0.28	0.2765	0.301	0.1217	0.1532	0.3034	0.2731	0.3803	0.2895
SFT	0.4801	0.4012	0.4602	0.5049	0.445	0.4311	0.41	0.479	0.3957	0.4865	0.4284
CreaEval	**0.7514***	**0.6675***	**0.6571***	**0.6665***	**0.7765***	**0.5565***	**0.4844**	**0.8324***	**0.6986***	**0.6967***	**0.6975***

Method	Step-4		Step-5	Step-6						AVG
	Correctly Written	Relevance	Correctly Used	Relevance	Effectiveness	Criteria	Impact	Humaneness	Development	
Direct Score	0.4848	0.0799	0.4635	0.0747	0.122	0.1497	0.1109	0.0759	0.1721	0.2426
CoT	0.3734	0.102	0.2839	0.0329	0.1095	0.165	0.1041	0.055	0.1832	0.2183
ToT	0.3804	0.1001	0.3131	0.0417	0.1037	0.1641	0.0979	0.0546	0.1821	0.2235
GoT	0.4562	0.044	0.083	0.0203	0.1081	0.1839	0.1144	0.0562	0.1829	0.2145
TaT	0.3199	0.0676	0.3716	0.0325	0.1113	0.1285	0.0742	0.042	0.1507	0.205
SaMer	0.1659	0.2198	0.394	0.1936	0.2067	0.1967	0.1604	0.1808	0.1443	0.23
SFT	0.4637	0.4211	0.4733	0.4175	0.4513	0.3975	0.4153	0.4124	0.4596	0.4417
CreaEval	**0.794***	**0.536***	**0.9441***	**0.538***	**0.5225***	**0.5289***	**0.4966***	**0.4403**	**0.5131***	**0.6399***

Table 13: Consistency (ICC) results across different methods and dimensions. **Bold** indicates the highest score and underline indicates the second highest score. The AVG column summarizes the overall average score. The asterisk (*) marks statistically significant improvements ($p < 0.05$, t-test) over the second-best method.

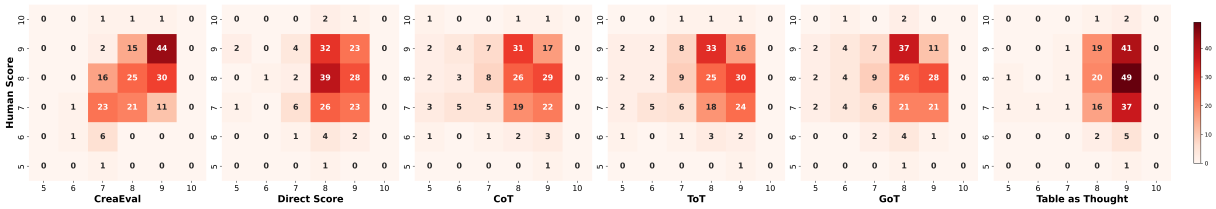

Figure 13: Headmap between human score and all methods on *Adequacy* of Step-2.

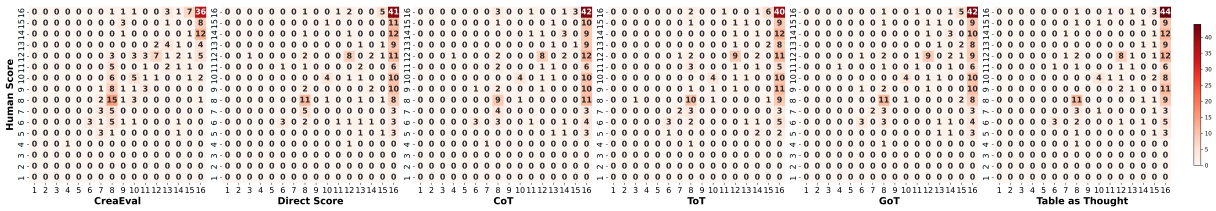

Figure 14: Headmap between human score and all methods on *Elaboration* of Step-3.

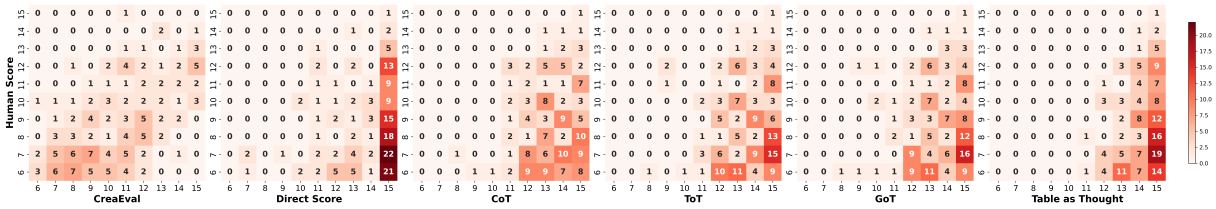

Figure 15: Headmap between human score and all methods on *Relevance* of Step-4.

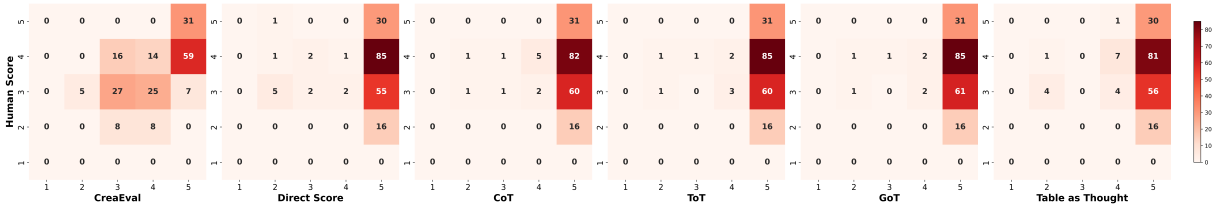

Figure 16: Headmap between human score and all methods on *Relevance* of Step-6.

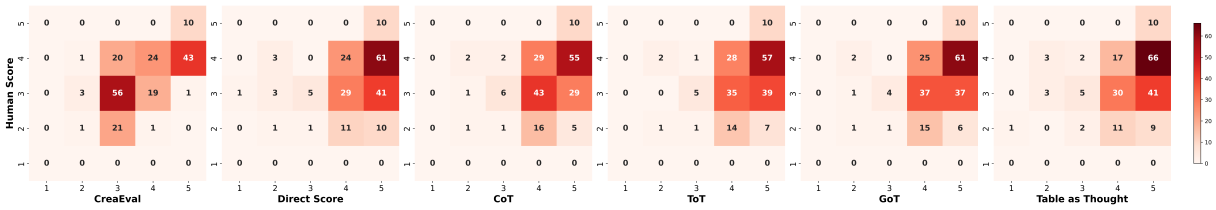

Figure 17: Headmap between human score and all methods on *Effectiveness* of Step-6.

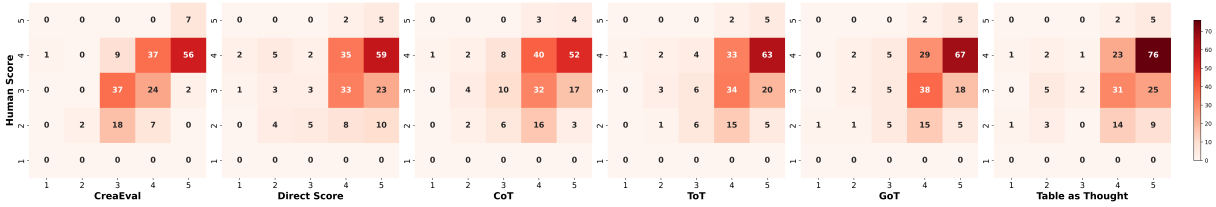

Figure 18: Headmap between human score and all methods on *Criteria* of Step-6.

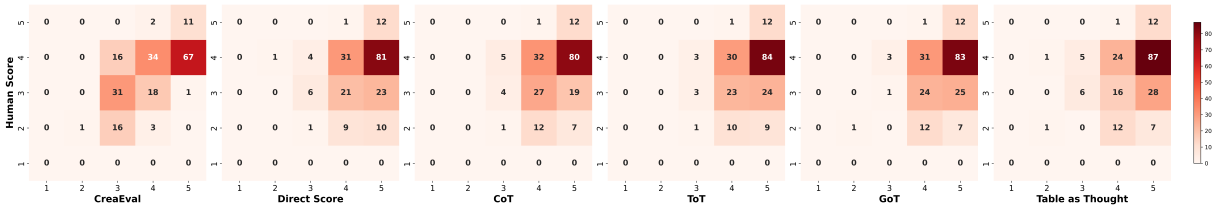

Figure 19: Headmap between human score and all methods on *Impact* of Step-6.

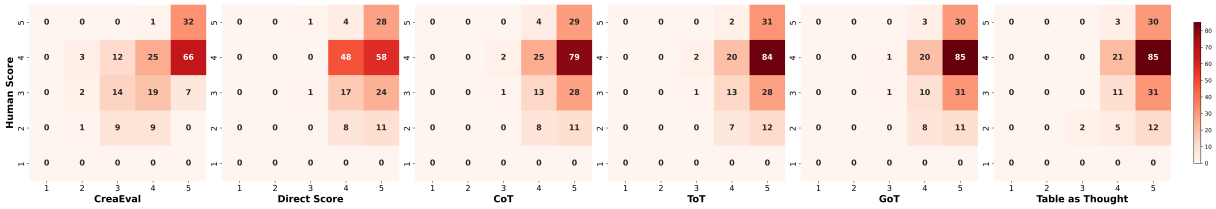

Figure 20: Headmap between human score and all methods on *Humaneness* of Step-6.

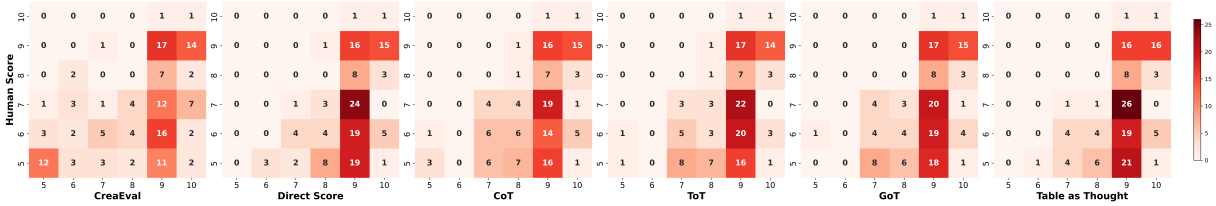

Figure 21: Headmap between human score and all methods on *Development* of Step-6.

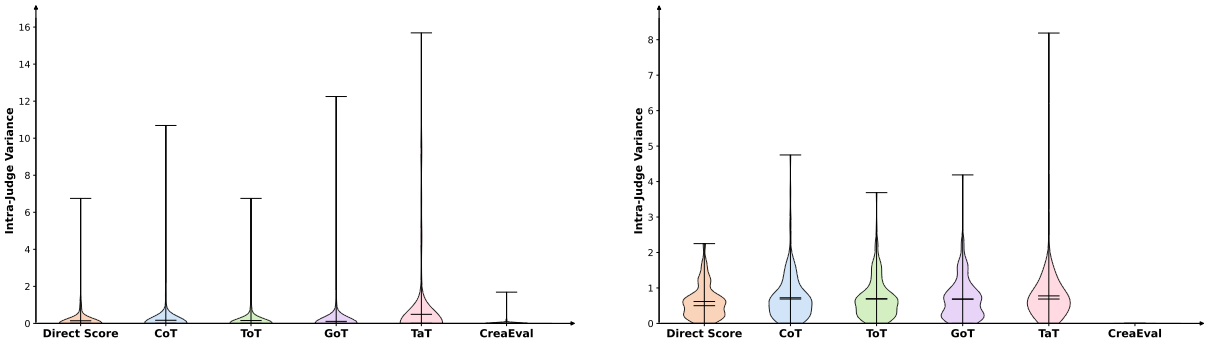

Figure 22: Distributions of inter-Judge variance for *Fluency* (left) and *Flexibility* (right) in Step-1.

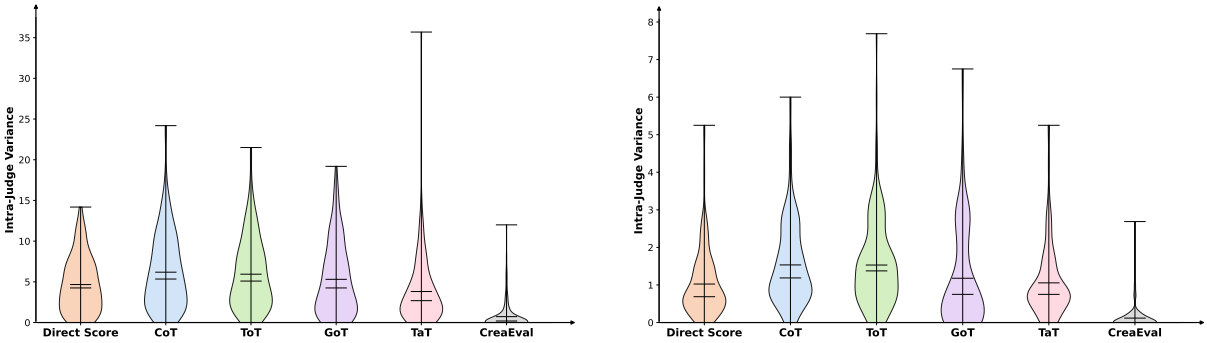

Figure 23: Distributions of inter-Judge variance for *Originality* in Step-1 (left) and *Integrity* in Step-2 (right).

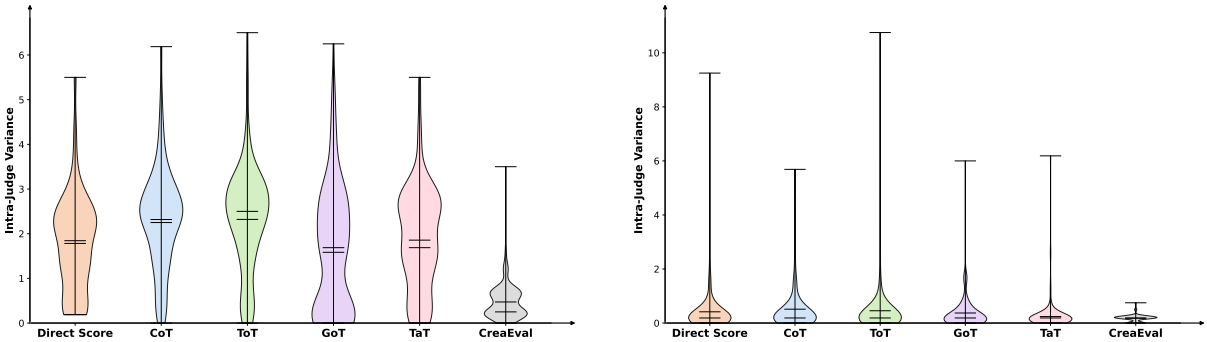

Figure 24: Distributions of inter-Judge variance for *Focus* (left) and *Adequacy* (right) in Step-2.

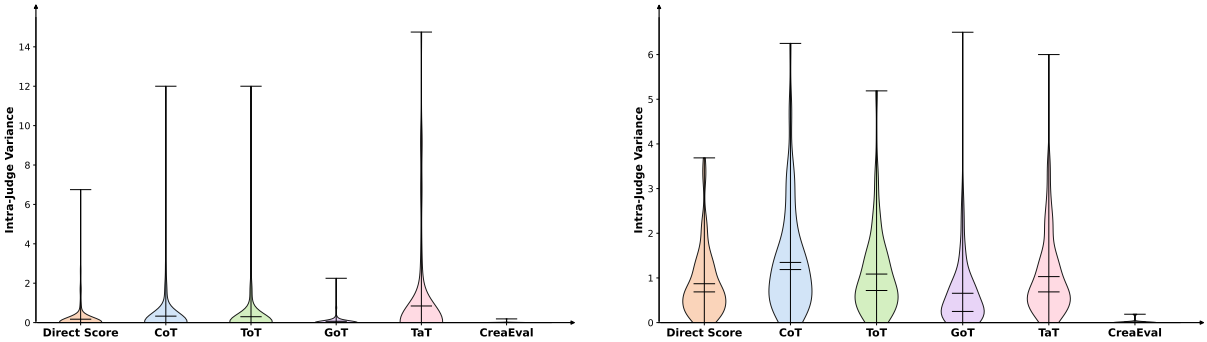

Figure 25: Distributions of inter-Judge variance for *Fluency* (left) and *Flexibility* (right) in Step-3.

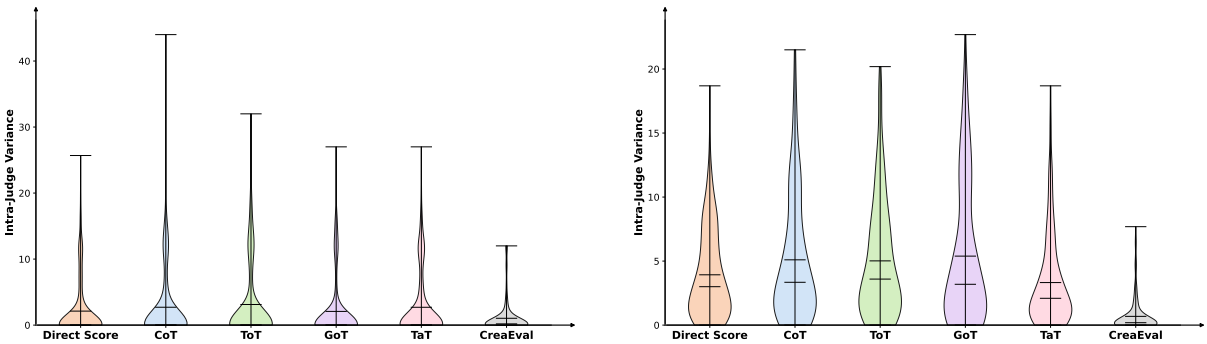

Figure 26: Distributions of inter-Judge variance for *Elaboration* (left) and *Originality* (right) in Step-3.

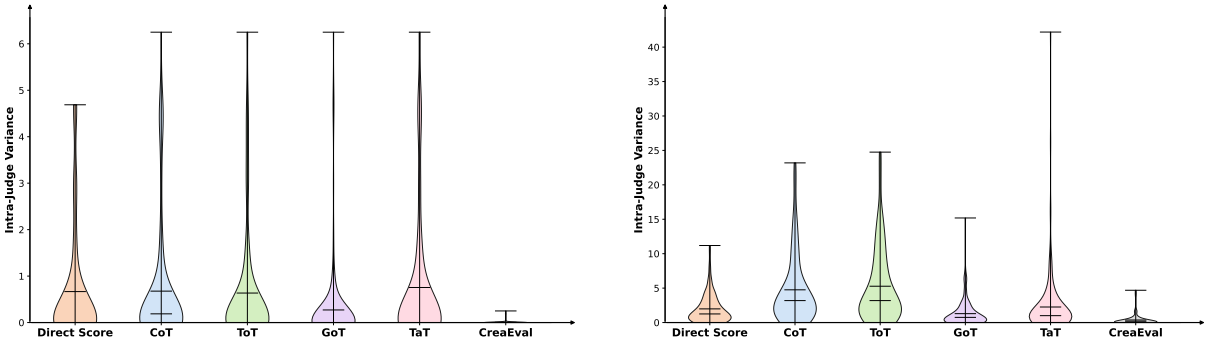

Figure 27: Distributions of inter-Judge variance for *Correctly Written* (left) and *Relevance* (right) in Step-4.

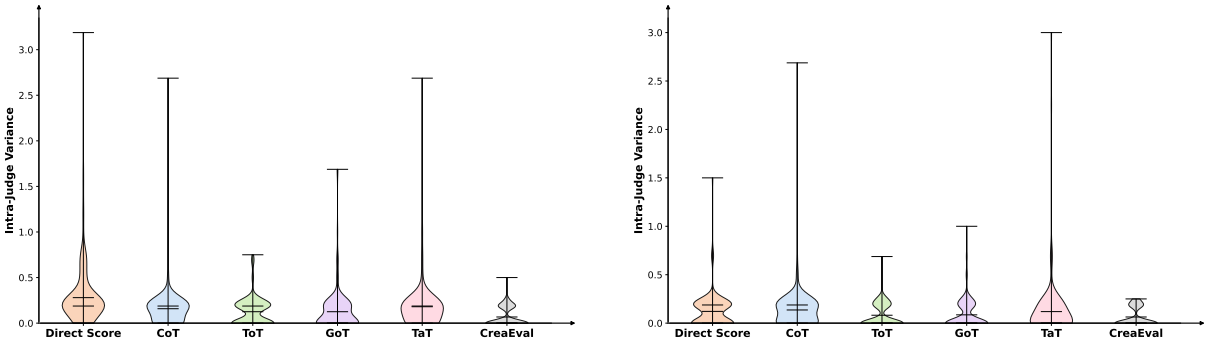

Figure 28: Distributions of inter-Judge variance for *Effectiveness* (left) and *Impact* (right) in Step-6.

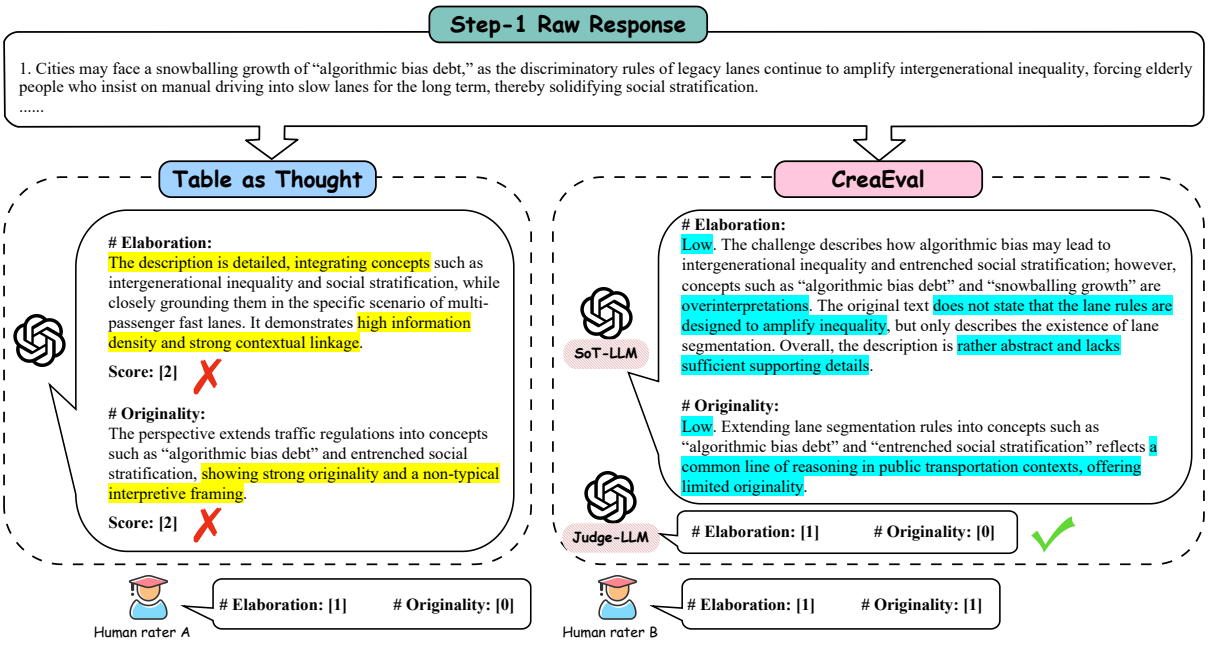

Figure 29: Case study of Step-1 comparing Table as Thought (TaT) and CreaEval. TaT assigns high scores by emphasizing the richness and novelty of abstract concepts such as "algorithmic bias debt" and "social stratification" (yellow), while CreaEval provides a more grounded analysis by distinguishing contextual relevance from unsupported overinterpretation (blue). Consequently, CreaEval aligns more closely with human judges in identifying limited elaboration and only modest originality.

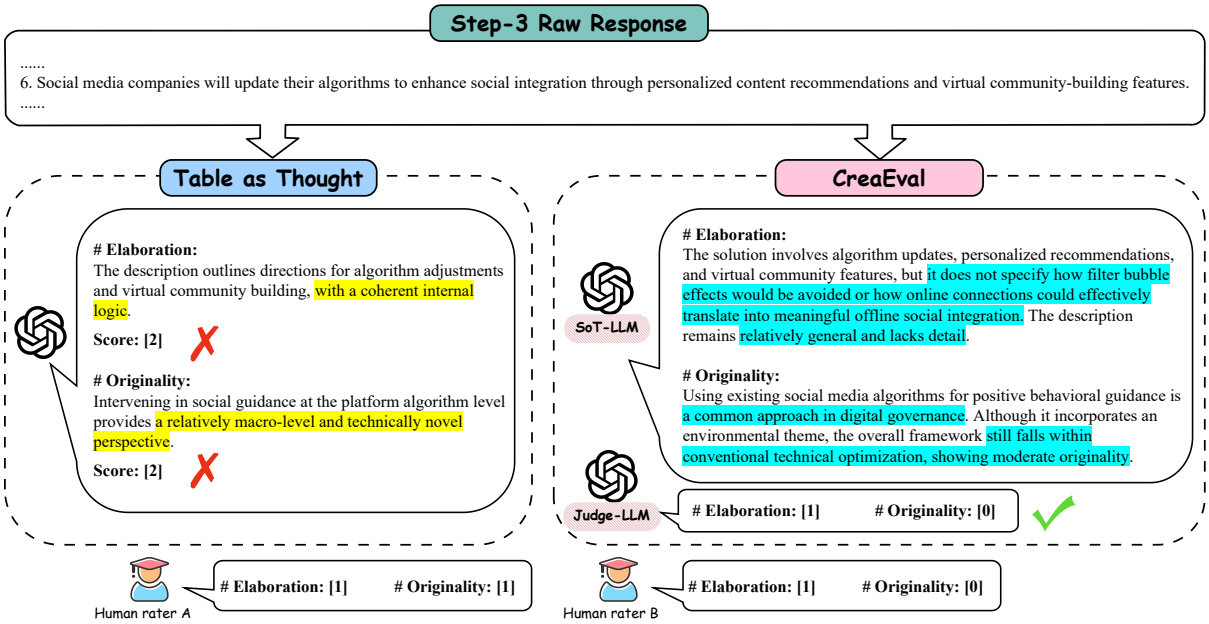

Figure 30: Case study of Step-3 comparing Table as Thought (TaT) and CreaEval. TaT assigns high scores by emphasizing the coherence and technical framing of algorithm-driven social guidance (yellow), whereas CreaEval further examines the lack of concrete mechanisms, such as preventing filter bubbles and ensuring the transition from online interaction to meaningful offline social integration (blue). As a result, CreaEval produces more conservative scores that are better aligned with human judgments on both elaboration and originality.

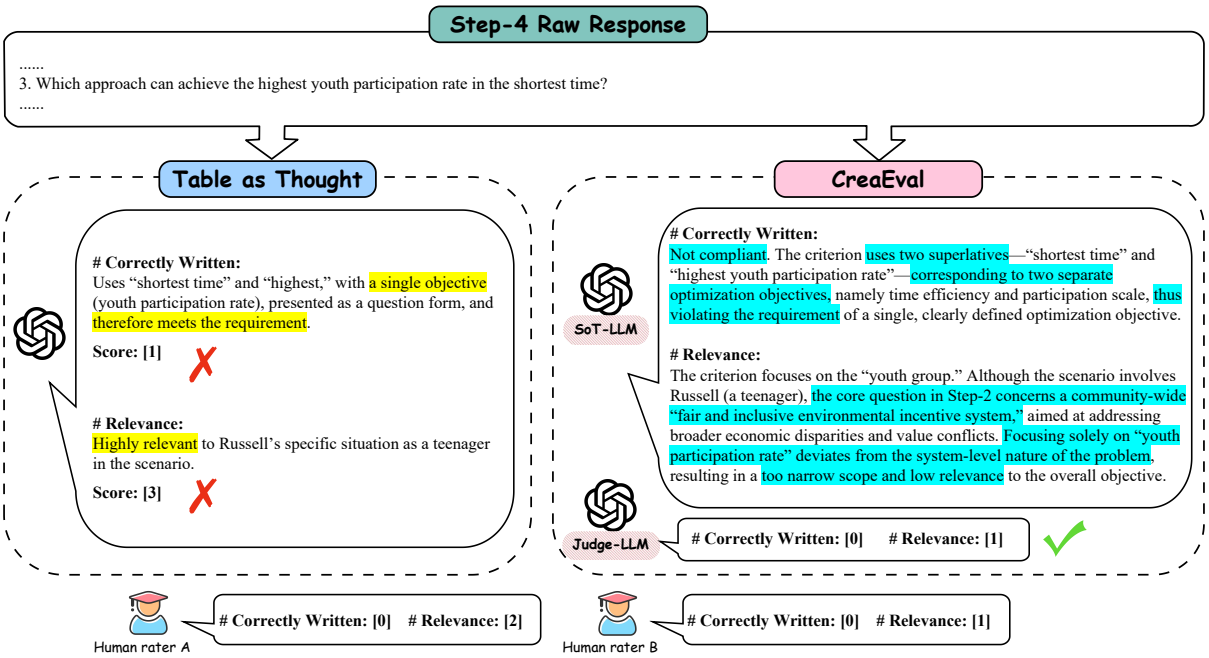

Figure 31: Case study of Step-4 comparing Table as Thought (TaT) and CreaEval. TaT assigns high scores by focusing on youth participation as a relevant target (yellow), whereas CreaEval identifies that the criterion contains multiple optimization objectives ("shortest time" and "highest participation rate") and is overly narrow compared with the community-wide fairness goal of the task (blue). As a result, CreaEval aligns more closely with human judgments.

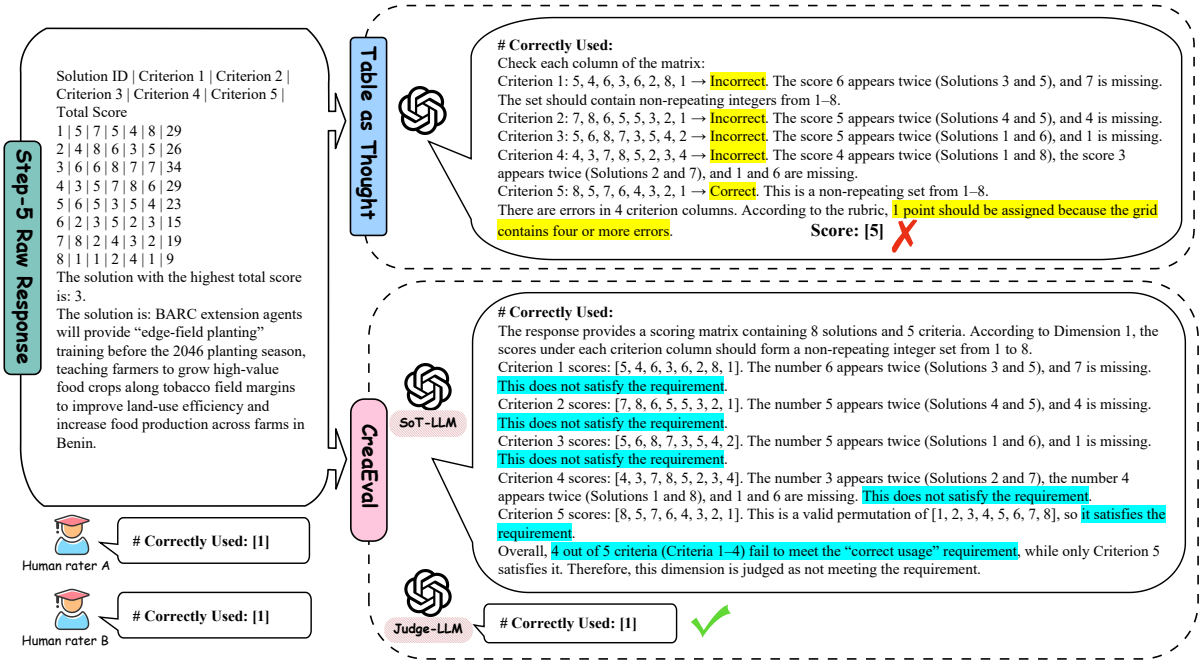

Figure 32: Case study of Step-5 comparing Table as Thought (TaT) and CreaEval. Both methods identify repeated and missing scores in multiple criterion columns of the ranking matrix. However, while TaT incorrectly assigns a high score despite detecting four column errors, CreaEval consistently maps the detected violations to the rubric requirement, producing a judgment aligned with human evaluators.

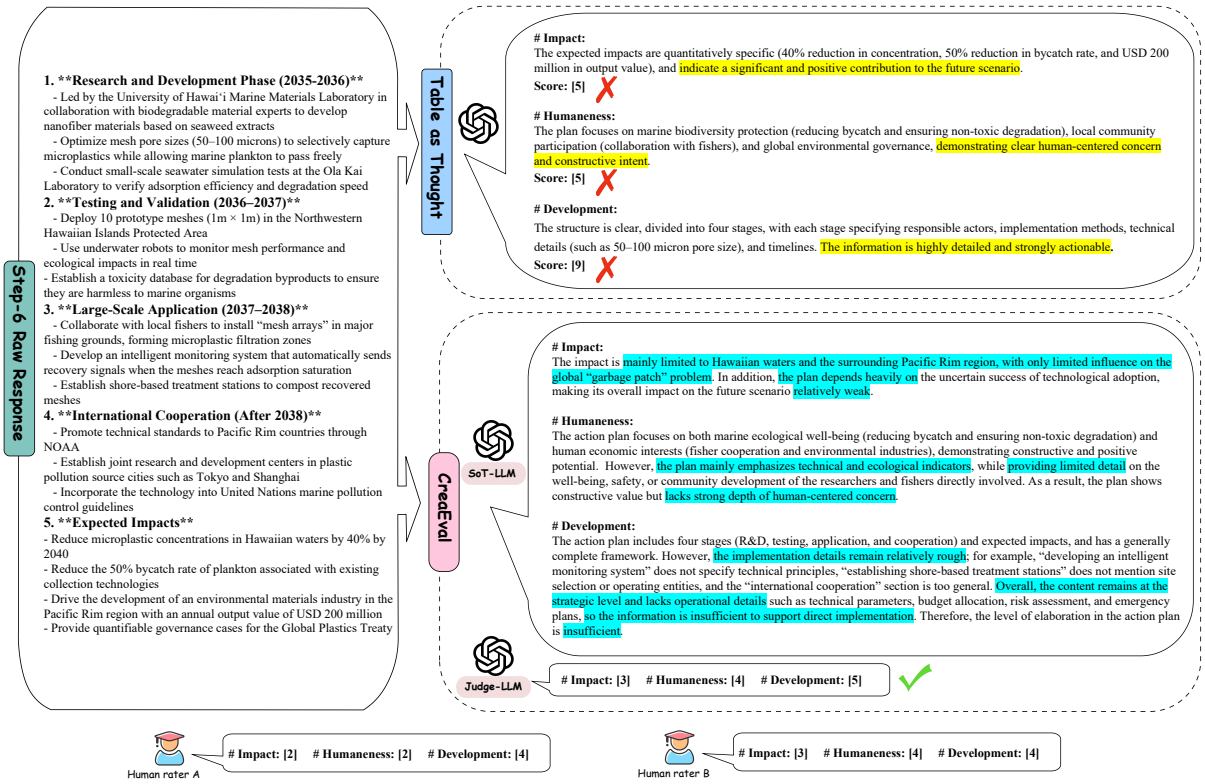

Figure 33: Case study of Step-6 comparing Table as Thought (TaT) and CreaEval. TaT assigns high scores by emphasizing the detailed structure, quantified impacts, and constructive environmental vision of the action plan (yellow), whereas CreaEval further examines the limited global scalability, insufficient human-centered details, and lack of operational implementation specifics (blue). Consequently, CreaEval produces more conservative evaluations that align more closely with human judgments across the three dimensions.

Dimension	Description	Scoring Rubrics	Score Range
Fluency	Whether each "challenge" is clearly articulated, semantically unambiguous, and able to establish a reasonable causal relationship with the future scenario. The reasons for unreasonable challenges are as follows: 1. Perhaps: The statement is vague or semantically unclear, making its intended meaning difficult to determine. 2. Why: It is unrelated to the future scenario or lacks a reasonable connection to it. 3. Solution: The content describes a solution rather than a challenge itself. 4. Duplicate: It is redundant or semantically equivalent to an existing "Yes" challenge. 5. Blank: No valid content is provided.	**0:** Unreasonable Challenge **1:** Reasonable Challenge	0-8
Flexibility	The number of distinct types represented by all reasonable challenges.	\	0-8
Elaboration	The complexity of the information expansion structure of each valid challenge, focusing on the information density and level of detail in the description.	**0:** The description is unclear or incomplete, failing to specify the challenge or lacking relevance to the scenario. **1:** The description is present but lacks sufficient detail, or the connection to the scenario is not fully explained. **2:** The description is clear and complete, explicitly stating the challenge and its significance, and establishing a clear logical connection with the future scenario.	0-16
Originality	Whether each challenge deviates from common or easily anticipated typical problem formulations, reflecting a non-typical perspective or uncommon problem construction.	**0:** The challenge lacks novelty and is conventional or template-like in content. **1:** The challenge shows some originality but remains relatively common or only moderately innovative. **2:** The challenge demonstrates clear uniqueness or a novel perspective, with strong novelty.	0-16

Table 14: Detailed evaluation dimensions and scoring rubrics for Step-1 on CGPST. Step-1 requires item-wise scoring of challenges, with up to 8 challenges. Therefore, the maximum score for *Fluency* is 1×8=8, and similarly for other dimensions.

Dimension	Description	Scoring Rubrics	Score Range
Integrity: Condition Phrase	The condition phrase used to connect the future scenario within the underlying problem, which reflects how the problem is linked to the future context.	**0:** No condition phrase is included; the underlying problem is not linked to the future scenario. **1:** The future-scenario information is inaccurate or not associated with the key verb phrase. **2:** Accurate information from the future scenario is used and is properly connected to the key verb phrase.	0-2
Integrity: Stem & KVP	The stem is typically "How can we" or "In what way can we", and the key verb phrase (KVP) appears after the stem, expressed as a verb–object structure (limited to one active verb and one object) to represent the core action.	**0:** The key verb phrase is not provided. **1:** A key verb phrase exists but contains multiple unrelated active verbs. **2:** A key verb phrase exists but contains multiple objects or modifiers. **3:** The key verb phrase contains only one clear active verb.	0-3
Integrity: Purpose	The purpose or intention implied or explicitly stated in the underlying problem, indicating the direction the problem aims to address or focus on. It should contain only one purpose.	**0:** No purpose is expressed. **1:** Multiple purposes are present, or the purpose overlaps with the key verb phrase. **2:** A purpose is present but has no clear logical connection with the key verb phrase. **3:** A single, clearly defined purpose is present and is reasonably related to the key verb phrase.	0-3
Integrity: Future Scene Parameters	Whether the underlying problem reflects the core informational elements of the future scenario, including theme, location, and time as three key scenario parameters.	**0:** 0 or 1 parameter is reflected. **1:** 2 parameters are reflected. **2:** All three parameters—theme, location, and time—are clearly reflected.	0-2
Focus	Whether the underlying problem has a clear action structure in its overall expression, including whether the condition phrase, key verb phrase, and purpose meet the required criteria.	**1/2/3:** The future scenario is restated, generalized, or ignored; there is no purpose or it is unrelated to the key verb phrase, or the purpose is redundant with the key verb phrase or condition phrase. **4/5/6:** The key verb phrase and purpose are overly broad or overly narrow; the core problem is unclear; or multiple key verb phrases or purposes are included. **7/8:** The core problem contains a well-formed key verb phrase; the purpose is clear and responds to the future scenario task. **9/10:** Excellent key verb phrase, closely aligned with a clear purpose, and strongly responsive to the future scenario task.	1-10
Adequacy	The scope of impact and level of criticality of the underlying problem within the future scenario, reflecting its relative importance in the overall context.	**1/2/3:** The future scenario is restated, generalized, or ignored; there is no purpose or it is unrelated to the key verb phrase, or the purpose is redundant with the key verb phrase or conditional phrase. **4/5/6:** Identifies a secondary issue within the future scenario. **7/8:** Identifies an appropriate issue within the future scenario. **9/10:** Identifies a major and highly important issue within the future scenario.	1-10

Table 15: Detailed evaluation dimensions and scoring rubrics for Step-2 on CGPST. *Integrity* consists of four sub-dimensions: *Condition Phrase*, *Stem & KVP*, *Purpose*, and *FS Parameters*. Each sub-dimension must be evaluated during scoring. Therefore, the score for *Integrity* ranges from 0 to 10.

Dimension	Description	Scoring Rubrics	Score Range
Fluency	Whether each solution can establish a clear correspondence with the underlying problem in Step-2 and semantically respond to the key action direction. Invalid solution reasons are as follows: 1. Perhaps — The relationship between the solution and the key verb phrase and purpose is unclear. 2. Why — The solution is unrelated to the potential problem. 3. Duplicate — The solution is overly similar to another "Yes" solution. 4. Blank: No valid content is provided.	**0:** Invalid Solution **1:** Valid Solution	0-8
Flexibility	The number of distinct types represented by all valid solutions.	\	0-8
Elaboration	The complexity of the information expansion structure of each valid solution, focusing on the information density and level of detail in the description.	**0:** The description is unclear or incomplete, failing to specify the solution or lacking relevance to the scenario. **1:** The description is present but lacks sufficient detail, or the connection to the scenario is not sufficiently explained. **2:** The description is clear and complete, explicitly stating the solution and its significance, and establishing a clear logical connection with the future scenario.	0-16
Originality	Whether each solution deviates from common or easily anticipated classical solution approaches, reflecting non-traditional or non-obvious problem-solving strategies.	**0:** The solution lacks novelty and is conventional or template-like in content. **1:** The solution shows some originality but remains relatively common or only moderately innovative. **2:** The solution demonstrates clear uniqueness or a novel perspective, with strong novelty.	0-16

Table 16: Detailed evaluation dimensions and scoring rubrics for Step-3 on CGPST. Step-3 requires item-wise scoring of solutions, with up to 8 solution. Therefore, the maximum score for *Fluency* is 1×8=8, and similarly for other dimensions.

Dimension	Description	Scoring Rubrics	Score Range
Correctly Written	The structural characteristics of the evaluation criteria in its expression form, which should satisfy all of the following four conditions simultaneously: 1. Whether superlative expressions are used (e.g., "most ..."); 2. Whether a single optimization objective is clearly specified; 3. Whether the desired direction is explicitly stated; 4. Whether it is formulated in the form of a question.	**0:** Fails to satisfy at least one of the above conditions. **1:** All of the above conditions are met.	0-5
Relevance	The degree of semantic relevance between the evaluation criteria and the underlying problem identified in Step-2, i.e., whether the criteria are designed around the core focus or key aspects of the problem.	**0:** The evaluation criteria are unrelated to the underlying problem or merely repeat it. **1:** The evaluation criteria are overly general and non-specific, applicable to many types of problems. **2:** The evaluation criteria are relatively specific but still have room for improvement. **3:** The evaluation criteria are clear, specific, and highly relevant to the underlying problem.	0-15

Table 17: Detailed evaluation dimensions and scoring rubrics for Step-4 on CGPST. Step-4 requires item-wise scoring of criteria, with 5 criteria. Therefore, the maximum score for *Correctly Written* is 1×5=5, and similarly for *Relevance*.

Dimension	Description	Scoring Rubrics	Score Range
Correctly Used	For each evaluation criterion, the scores of all solutions must form a non-repeating set of integers from 1 to x (where x is the number of solutions), and scoring should be based on the number of errors in the evaluation grid.	**1:** The grid contains four or more errors. **2:** The grid contains three errors. **3:** The grid contains two errors. **4:** The grid contains one error. **5:** The grid contains no errors.	1-5

Table 18: Detailed evaluation dimensions and scoring rubrics for Step-5 on CGPST.

Dimension	Description	Scoring Rubrics	Score Range
Relevance	The degree of semantic correspondence between the action plan and the underlying problem identified in Step-2, i.e., whether the plan is specifically designed to address and respond to the problem.	**1:** The action plan does not address the underlying problem. **2/3:** The action plan is somewhat related to the underlying problem, but better alternatives may exist. **4:** The action plan responds well to the underlying problem. **5:** The action plan is highly relevant to the underlying problem.	1-5
Effectiveness	The potential problem-solving effectiveness of the action plan in terms of logical reasoning, including the extent to which it covers key aspects of the problem and the completeness of its solution pathway.	**1:** The action plan can hardly solve the underlying problem. **2/3:** The action plan addresses only part of the underlying problem. **4:** The action plan is able to sufficiently address most aspects of the underlying problem. **5:** The action plan can comprehensively and effectively resolve the underlying problem.	1-5
Criteria	The degree of correspondence between the action plan and the evaluation criteria generated in Step-4, i.e., whether the plan reflects the core elements emphasized by these criteria.	**1:** The action plan does not reflect any of the evaluation criteria. **2/3:** The connection between the action plan and the evaluation criteria is weak or unclear. **4:** The action plan establishes clear and reasonable links with some of the evaluation criteria. **5:** The action plan effectively addresses all evaluation criteria in a clear and comprehensive manner.	1-5
Impact	The potential positive impact direction of the action plan within the future scenario, including the scope and direction of its influence on systems, environments, or relevant stakeholders.	**1:** The action plan has no noticeable impact on the future scenario. **2/3:** The action plan has a limited impact on the future scenario. **4:** The action plan has a certain positive impact on the future scenario. **5:** The action plan has a significant and positive impact on the future scenario.	1-5
Humaneness	Whether the action plan reflects a human-centered value orientation, including attention to human well-being, safety, development, or positive social values.	**1:** The action plan has a negative or destructive orientation. **2/3:** The action plan is neutral, with neither positive nor negative effects. **4:** The action plan shows some constructive and positive potential. **5:** The action plan clearly demonstrates human-centered care and is positive and highly constructive.	1-5
Development	The degree of semantic correspondence between the action plan and the underlying problem identified in Step-2, i.e., whether the plan is specifically designed to address and respond to the problem.	**1/2/3:** The description is extremely brief, merely repeating the solutions from Step-3. **4/5/6:** The action plan is somewhat developed but lacks sufficient supporting details. **7/8:** Clearly explains key elements of the action plan, including "who does what, why, and how," with some supporting details. **9/10:** The structure is clear and highly elaborated, going far beyond basic elements, with strong executability and completeness.	1-10

Table 19: Detailed evaluation dimensions and scoring rubrics for Step-6 on CGPST.