

AlphaContext: An Evolutionary Tree-based Psychometric Context Generator for Creativity Assessment

Yixuan Wang^{1*} Yue Huang^{1*} Hong Qian^{1,2†} Yunzhao Wei¹ Yifei Ding¹
Wenkai Wang¹ Zhi Liu^{1,2} Zhongjing Huang¹ Aimin Zhou^{1,2} Jiajun Guo¹

¹East China Normal University, Shanghai, China

²Shanghai Innovation Institute, Shanghai, China

{yxiwang, yhuang}@stu.ecnu.edu.cn, hqian@cs.ecnu.edu.cn

Abstract

Creativity has become a core competence in the era of LLMs and human–AI collaboration, underpinning innovation in real-world problem solving. Crucially, the systematic improvement of creativity necessitates scientifically valid assessment instruments. Psychometric research recognizes context-based assessment as an effective way to measure creative thinking. However, high-quality expert-designed contexts remain scarce. Existing LLM-based generators often struggle with insufficient assessment cues, weak narrative coherence, limited stylistic diversity, and poor support for creative thinking. To address these challenges, we propose AlphaContext, an evolutionary tree-based psychometric context generator for creativity assessment. First, the HyperTree Outline Planner formalizes expert-designed outlining as a rule-guided hypertree and performs top-down hierarchical planning. The MCTS-based Context Generator fills the outline via MCTS to balance global structure and local quality. Then, the Evolutionary Context Optimizer evolves contexts with MAP-Elites by repeatedly updating niche elites to jointly improve diversity and quality. Finally, the Assessment-Guided Evolution Refiner simulates virtual participants with diverse styles and recycles weak contexts for further evolution. Experiments show that AlphaContext yields an average improvement of 8% over competitive methods across 6 quality metrics.

1 Introduction

Creativity is typically defined as the ability to generate novel and appropriate ideas (Runco and Jaeger, 2012), and it is a crucial skill that drives social innovation and scientific discovery (Sternberg and Lubart, 1999). As AI increasingly takes over routine tasks, creativity is becoming an even more important driver of original contributions and transformative breakthroughs (Lee, 2022).

*Equal Contribution

†Corresponding Author

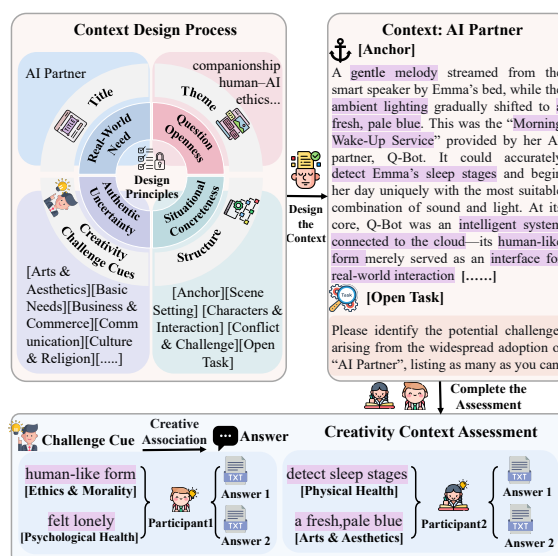


Figure 1: Workflow of creativity context assessment. Experts first design an “AI Partner” psychometric context containing implicit challenge cues. Participants then complete an open-ended task to identify potential challenges based on the context in their responses.

Given this growing importance, scientifically assessing creativity has become a key issue in psychometrics and intelligent education (Jr. et al., 2024; Wang et al., 2025b; Guo and Woulfin, 2016). In creativity assessment, context-based paradigms have been widely adopted (Barbot et al., 2019b). Sternberg’s triarchic theory of intelligence emphasizes that creative thinking arises when faced with novel situations (Sternberg, 1984). Therefore, future-oriented contexts, due to their inherent high uncertainty and broad imaginative space, are considered ideal stimuli for eliciting creativity thinking (Koh and Leung, 2019), as shown in Figure 1. The Future Problem Solving Program (FPSP) (Torrance et al., 1976) provides authoritative evidence that well-designed future contexts can reliably elicit creative thinking (Crabbe, 1982). Therefore, generating high-quality contexts is essential for developing valid and reliable creativity assessments

for humans. However, current practice faces a significant bottleneck in productivity. High-quality creativity assessment contexts still rely on expert craftsmanship.

In recent years, rapid advances in LLMs have substantially improved the generation of stories and dialogues (Li et al., 2025a; Feng et al., 2025; Wu et al., 2025a), making the automated construction of psychometric contexts for the assessment of creativity increasingly plausible. Although much prior work has examined the creativity of LLMs, there has been far less research on whether LLMs can generate valid creativity assessment contexts for humans. However, psychometric contexts differ from general narratives, and both directly influencing LLMs and reusing story-generation frameworks still face two key challenges. *The first key challenge is to simultaneously enforce implicit assessment cues and global narrative coherence.* Psychometric cues for creative thinking are often embedded implicitly in textual details. Existing methods struggle to precisely control the alignment between cues and themes (Bai et al., 2025), thus failing to satisfy the fine-grained constraints required for psychometric content design and narrative structuring. *The second key challenge is to improve diversity while ensuring both context quality and measurement validity at limited cost.* For a given theme, future problem contexts require diverse types and styles to adapt to different assessment populations (Novikov et al., 2025), yet increasing diversity typically raises generation and refinement costs. Moreover, creativity assessment contexts require reliable quality and elicitation validity. Traditional expert workflows as shown in Figure 1, rely on expensive human studies and iterative rework (Wu et al., 2025b), while current methods lack efficient validation and optimization loops. Our work has implications for employing LLMs to automatically generate valid and reliable creativity assessment contexts for humans and AI.

To address these two challenges, this paper proposes AlphaContext, an evolutionary tree-based psychometric context generator for creativity assessment. To tackle the first challenge, the HyperTree Outline Planner formalizes context outlining as a rule-guided hypertree, mapping expert reasoning into a searchable outline space. The MCTS-based Context Generator then performs Monte Carlo Tree Search (MCTS) generation under the outline, balancing global structural coherence and local semantic quality to produce seed contexts. To

handle the second challenge, the Evolutionary Context Optimizer conducts evolutionary search with MAP-Elites in a task-specific behavioral space, iteratively expanding stylistic diversity via niche-wise elite updates. Finally, the Assessment-Guided Evolution Refiner simulates virtual participant responses and iteratively refines weak contexts to better elicit creative thinking. Experimental results show that AlphaContext substantially outperforms baselines across multiple evaluation metrics.

2 Related Work

2.1 Language Model based Creativity Assessment

The NLP community has shown growing interest in psychometrically grounded creativity assessment. Luchini et al. (Luchini et al., 2025) fine-tuned RoBERTa and GPT-2 to automatically score creativity responses. Since assessment contexts are expert-designed and costly to scale, CPIG investigates using LLMs to automatically generate items for a classic free-response creativity test, examining whether LLMs can generate valid creativity assessments for humans. However, CPIG targets short items and does not address long-form Future Problem contexts, which require discourse-level coherence and implicit assessment cues.

In parallel, other works evaluate the creativity of LLMs (Si et al., 2025; Fang et al., 2025). LiveIdeaBench (Ruan et al., 2024) uses single-keyword prompts to assess scientific creative thinking. AidanBench (Mclaughlin et al., 2024) measures novelty, non-redundancy, and coherence under open-ended creative questions. Nevertheless, these benchmarks rely on keyword triggers or expert-designed contexts. Moreover, the scarcity of high-quality long-form contexts limits both evaluation protocols and model improvement.

Overall, the key bottleneck is generating high-quality long-form psychometric contexts that enable valid creativity assessment for humans and also support more comparable evaluation of LLM creativity. To address this gap, we propose AlphaContext, an evolutionary tree-based psychometric context generator for creativity assessment.

2.2 LLM-based Story Generation

LLMs have significantly advanced automated narrative generation (Bai et al., 2025; Wu et al., 2025a; Lee et al., 2025), enabling the generation of coherent long-form stories. DOC (Yang et al., 2023)

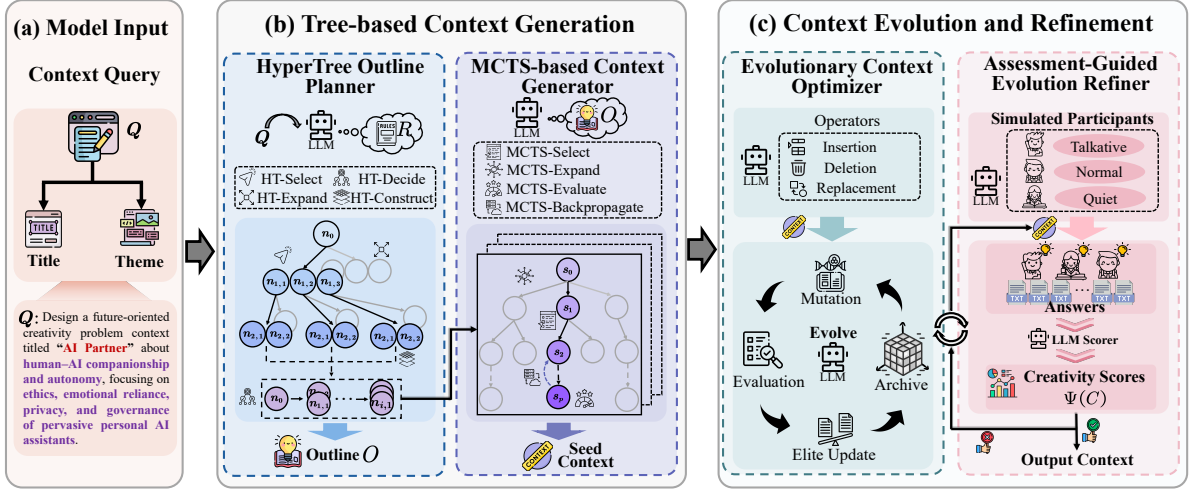


Figure 2: The procedure of the proposed AlphaContext. (a) Given a context query Q , (b) the HyperTree Outline Planner and MCTS-based Context Generator generate seed contexts, and (c) the Evolutionary Context Optimizer and Assessment-guided Evolution Refiner improve diversity and quality, yielding assessment-ready contexts.

adopts the outline-first strategy and then expands the outline into detailed text. STORYTELLER (Li et al., 2025a) introduces a plot node mechanism based on the subject-verb-object (SVO) structure and a dynamic interaction module, further improving narrative coherence and logical consistency. However, most existing story generation methods typically focus on entertainment and fluency, often failing to satisfy the quality and validity requirements of psychometric assessment contexts. Although SS-GEN (Feng et al., 2025) explores generating psychological social stories for autism interventions, this setting differs fundamentally from creativity assessment settings. Creativity assessment contexts must maintain coherent long-form narratives while implicitly placing assessment cues that elicit creative thinking and support psychometric validity. Consequently, generating high-quality long-form psychometric contexts for creativity assessment remains an open challenge.

3 Preliminaries

Creativity Context Generation. Given a context design query Q that specifies the title and theme, together with a pre-trained LLM π_θ , our goal is to construct an assessment-ready archive $\mathcal{A} = \{C_k\}_{k=1}^{|\mathcal{A}|}$ of contexts for creativity assessment. Each context is represented as $C = (T, O)$, where O is a structured outline, and T is the resulting context text guided by O . We adopt a plan-generate-evolve pipeline. The planner produces $O = \pi_\theta(\Phi(Q))$ under predefined instructions Φ . The generator realizes T guided by the outline O .

The evolve stage iteratively refines and diversifies candidates by updating \mathcal{A} .

HyperTree Structure. In conventional tree-based planning, each edge links a parent node to a single child node. By contrast, a HyperTree introduces directed hyperedges, where a parent node connects to a set of child nodes via one edge, enabling hierarchical divide-and-conquer by jointly organizing discourse structure and assessment-cue placement for outline planning. Formally, we define a HyperTree as $\mathcal{H} = (N, Q, \mathcal{R})$, where Q denotes the query, N is the node set, and \mathcal{R} is a set of expansion rules. Given Q , the HyperTree is generated hierarchically according to the rule set \mathcal{R} . Compared with ordinary trees, this structure better aligns with expert practices in creativity context design.

4 The Proposed AlphaContext

Overview. As shown in Figure 2, AlphaContext comprises four modules: the HyperTree Outline Planner, the MCTS-based Context Generator, the Evolutionary Context Optimizer and the Assessment-Guided Evolution Refiner. Given a query Q , the HyperTree Outline Planner places expert outlining as a rule-guided hypertree over a library \mathcal{R} . The MCTS-based Context Generator then performs a sentence-level search to fill in the outline. To cover the multi-solution space under the same theme, the Evolutionary Context Optimizer applies MAP-Elites to explore diverse styles in a task-specific behavior space while improving within-niche quality. Finally, the Assessment-Guided Evolution Refiner simulates participant re-

sponses and feeds weak contexts back for further evolution.

4.1 HyperTree Outline Planner

Experts plan contexts holistically and refine them hierarchically, motivating a HyperTree representation. We propose the HyperTree Outline Planner to cast outline design as a HyperTree search, where directed hyperedges support hierarchical divide-and-conquer over structure and cue placement. Formally, we define a HyperTree as $\mathcal{H} = (N, Q, \mathcal{R})$. Given Q , HyperTree \mathcal{H} is generated hierarchically under \mathcal{R} . Each node $n \in N$ corresponds to a structural unit, and each rule $r \in \mathcal{R}$ expands a parent node into a set of child nodes through $r : n_p \mapsto n_c$, where n_p is the parent node and n_c denotes the corresponding child nodes. The HyperTree (HT) planner proceeds in four phases: HT-Select, HT-Expand, HT-Construct, and HT-Decide.

HT-Select. Given the current hypertree \mathcal{H} , its distinct branches are mapped onto a set of hyperchains $\{L_1, \dots, L_g\}$, where g denotes the number of hyperchains. To control the search scale, an LLM is used to evaluate and prune candidate hyperchains. Each candidate hyperchain is scored by the LLM and the optimal hyperchains \mathcal{L}^* are selected. The divisible nodes are then identified under the rule set \mathcal{R} . For each selected hyperchain, we choose its most promising divisible leaf node for expansion using an LLM selector. Thus, this phase consists of two steps: selecting the hyperchains \mathcal{L}^* and selecting the divisible leaf node n_i^* in each chosen hyperchain $L_i^* \in \mathcal{L}^*$.

HT-Expand. In this phase, given the selected node n_i^* in each hyperchain L_i^* , its applicable expansion rules are retrieved as $\mathcal{R}(n_i^*) = \{r \in \mathcal{R} \mid r : n_i^* \mapsto n_c\}$. For each rule $r \in \mathcal{R}(n_i^*)$, candidate child groups n_c are generated. Each group is treated as a single branch and is appended to L_i^* as the expansion outcome of n_i^* .

HT-Construct. The planner iterates Select and Expand step by step, growing \mathcal{H} from the root node across depth levels over a set of selected hyperchains. Construction stops when no divisible nodes remain or the iteration limit is reached, yielding a hypertree that compactly stores multiple candidate hyperchains.

HT-Decide. After constructing \mathcal{H} , the LLM globally evaluates the candidate hyperchains and decides the optimal hyperchain as the final outline O . This decision jointly considers structural validity for creativity assessment context design and nar-

rative consistency with the input title and theme. More details can be found in Appendix G.

4.2 MCTS-based Context Generator

In the generation stage, we cast creativity context writing as a sentence-level decision process guided by an outline O . Given an input prompt x and an outline O , we build a separate search tree for each discourse section. A context is represented as $C = (T, O)$, where $T = (t_1, \dots, t_p)$ is the generated sentence sequence. A node at depth p is $s_p = (t_p, N(s_p), V(s_p), O)$, where t_p is the current text, $N(s_p)$ is the visit count, and $V(s_p)$ is the estimated value. An LLM policy π_θ proposes the next candidate sentences, and an LLM evaluator provides quality feedback. The generator follows the standard MCTS loop: MCTS-Select, MCTS-Expand, MCTS-Evaluate, and MCTS-Backpropagate.

MCTS-Select. Starting from the root s_0 , the search recursively selects the child with the highest exploration potential according to the UCT score. Specifically, the UCT score of the node s_p is defined as $\text{UCT}(s_p) = V(s_p) + c\sqrt{\frac{\ln N(q)}{N(s_p)}}$. Here, $V(s_p)$ denotes the value score of s_p , $N(s_p)$ is its visit count, and $N(q)$ is the visit count of its parent node q . c is a hyper-parameter that balances exploitation ($V(s_p)$) and exploration (the second term).

MCTS-Expand. Given the selected node s_p , we expand it by sampling U candidate next sentences from the policy model $\pi_\theta: t_{p+1}^{(u)} \sim \pi_\theta(\cdot \mid x, t_{1:p}, O)$, $u = 1, \dots, U$. Here, $t_{1:p}$ denotes the previously generated sentences, enabling parallel exploration of diverse narrative realizations and cue instantiations within O .

MCTS-Evaluate. We evaluate each expanded node to assign its node value $V(s_{p+1})$. For long-form creativity context generation, we adopt a dual-horizon valuation mechanism at the evaluation phase to balance reliability and computational cost. Given an expanded child s_{p+1} , we first apply multi-aspect immediate scoring with an evaluator:

$$V_{\text{imm}}(s_{p+1}) = \bar{S}(s_{p+1})(1 - S_{\text{ha}}(s_{p+1})). \quad (1)$$

Here, $\bar{S}(s_{p+1})$ is a weighted average of cue alignment S_{sc} , imagery vividness S_{im} , and discourse coherence S_{co} with $\sum_i \omega_i = 1$. $S_{\text{ha}}(s_{p+1})$ measures hallucination risk. To mitigate myopic decisions, when $V_{\text{imm}}(s_{p+1}) < \tau$, we sample a short continuation and re-evaluate the concatenated fragment to obtain a more stable value estimate, a lightweight

look-ahead for this node. This look-ahead is triggered for low-scoring nodes to reduce myopic errors, while high-confidence nodes directly use the immediate evaluation to save sampling budget. For more details, please refer to our Appendix F.

MCTS-Backpropagate. The obtained evaluation score r_e is propagated back along the simulated path to all ancestor nodes s_j ($0 \leq j \leq p$), updating the visit counts and value estimates:

$$N_{\text{new}}(s_j) = N_{\text{old}}(s_j) + 1, \quad (2)$$

$$V_{\text{new}}(s_j) = \frac{V_{\text{old}}(s_j) N_{\text{old}}(s_j) + r_e}{N_{\text{new}}(s_j)}. \quad (3)$$

After multiple simulations, the tree concentrates on trajectories that better satisfy the outline, improve coherence, and reduce hallucination risk. We then extract the highest-value root-to-leaf path as a seed context to initialize the evolutionary module.

4.3 Evolutionary Context Optimizer

We introduce a MAP-Elites Evolutionary Context Optimizer initialized with the MCTS seed context. It maintains an elite archive in a style-oriented behavior space, expanding coverage and improving within-niche quality. We next describe the archive, mutation, evaluation, and update rules.

Diversity Archive. To characterize stylistic variations in creativity assessment contexts, we map each candidate context C into the behavior space B with a descriptor function $b(\cdot)$:

$$b(C) = [\phi_1(C), \phi_2(C), \phi_3(C)] \in [0, 1]^3. \quad (4)$$

Here, ϕ_1 captures proximity scope, measuring the extent to which a context is framed from personal daily-life settings to broader public issues. ϕ_2 captures knowledge density, reflecting how strongly the narrative is grounded in objective evidence such as data, mechanisms, and causal explanations. ϕ_3 captures viewpoint diversity, indicating the breadth of stakeholders involved and the need for multi-perspective integration. We uniformly discretize $[0, 1]^3$ to form a 3D grid archive, where each cell defines a behavioral niche and stores the current elite context with the highest fitness.

Mutation. In natural language space, we implement mutation as a conditional LLM editing policy π_θ that edits a parent elite context C_p . At each iteration, we apply an operator set $\Omega = \{\text{INSERTION}, \text{DELETION}, \text{REPLACEMENT}\}$ to revise key paragraphs and cue-bearing units,

introducing stylistic variation while maintaining assessment-critical content.

Evaluation. Given a mutated candidate C , we perform both behavior feature evaluation and quality evaluation. Feature evaluation computes $b(C)$ to determine the candidate’s niche assignment. Quality evaluation is produced by an LLM-based scorer in terms of narrative coherence S_{coh} , topical relevance S_{rel} , and engagement S_{eng} , and we define fitness $F(C) = \text{Avg}(S_{\text{coh}}(C) + S_{\text{rel}}(C) + S_{\text{eng}}(C))$. The scorer outputs three normalized quality scores for each candidate context, and the fitness is defined as their uniform average to avoid introducing extra hyperparameters.

Elite Update. After evaluation, we assign context C to the grid niche indexed by $b(C)$. If the niche is empty, C is inserted as the initial elite. Otherwise, let C^* denote the current elite in the niche; we replace C^* only when $F(C) > F(C^*)$.

4.4 Assessment-Guided Evolution Refiner

To ensure that generated creativity contexts reliably elicit measurable creative thinking, we propose an Assessment-Guided Evolution Refiner. Concretely, we instantiate an LLM-based participant simulator with a temperature set to 0 to suppress sampling randomness and stabilize response generation, making contexts comparable in a consistent simulation setting. To improve realism and interpretability, we model response styles as explicit profiles grounded in common psychometric response patterns and enforce each profile via role-conditioned prompting. We consider three profiles—*talkative*, *normal*, and *quiet*. Given a candidate context C , the simulator generates a set of responses $\{Y_m\}_{m=1}^M$. The refiner then scores each response using a creativity scorer $f_{\text{cre}}(\cdot)$. We define the assessment effectiveness of a context as the average creativity score across styles $\Psi(C) = \frac{1}{M} \sum_{m=1}^M f_{\text{cre}}(Y_m)$. If $\Psi(C)$ exceeds an expert-specified threshold, we treat C as an assessment-ready context. Otherwise, we route C back to the Evolutionary Context Optimizer for further iterative optimization.

5 Experiments

This section first describes the CreaTE dataset and details the evaluation metrics used. We then conduct extensive experiments to answer the following research questions. The codes and data are available at <https://github.com/yxwang19/AlphaContext>.

Table 1: Performance comparison across different methods on the CreaTE dataset. AlphaContext achieves the best results across all seven perspectives. All metrics are presented as positive percentages, where higher values indicate better performance. For each metric, the best-performing model is highlighted in **bold** and the second is underlined.

Methods	Coherence (↑)%	Relevance (↑)%	Engagement (↑)%	Significance (↑)%	Concreteness (↑)%	Uncertainty (↑)%	Diverse Verbs (↑)%
DeepSeek-V3.1	50.00	50.00	50.00	50.00	50.00	50.00	<u>94.33</u>
Qwen3-235B-A22B	41.50	47.91	45.07	33.87	46.55	53.33	92.09
Llama3.3-70B-Instruct	27.83	26.97	33.99	27.46	30.54	20.07	90.39
LongWriter-llama3.1-8b	26.60	27.46	28.63	23.40	33.62	25.99	91.18
LongWriter-glm4-9b	32.27	31.40	31.38	36.19	32.98	31.32	88.69
GPT-5.1	70.44	70.20	<u>65.39</u>	50.37	<u>71.80</u>	<u>68.60</u>	92.88
Gemini-3.0-Pro-Preview	<u>72.54</u>	<u>75.37</u>	62.56	48.40	64.16	63.30	91.81
DOC-v2	49.14	61.33	61.82	34.98	51.11	43.10	92.82
CRITICS	51.11	61.95	61.21	37.81	54.43	42.12	92.31
SS-GEN	60.22	69.69	56.40	<u>60.10</u>	51.85	53.57	90.24
AlphaContext	81.28	79.06	79.93	71.06	75.49	80.30	96.06

Q1: How does AlphaContext compare to existing methods in generating high-quality creativity contexts across multiple evaluation dimensions?

Q2: To what extent do the core components contribute to the performance of AlphaContext?

Q3: How does AlphaContext compare to other methods in terms of textual similarity with expert-designed contexts?

Q4: Does the LLM-based judge align with human preferences to support reliable evaluation?

Q5: Does AlphaContext remain effective for creativity assessment in real-world human studies?

Q6: How does the correlation of AlphaContext with expert-designed assessments compare to that of the strong LLM baseline?

Q7: What is the computational cost of AlphaContext in terms of generation time and token consumption?

5.1 Experimental Setup

Dataset. We evaluate AlphaContext on CreaTE. Since general story-generation datasets prioritize narrative completeness and style rather than assessment alignment, we construct *CreaTE*: 203 expert-curated title–theme inputs balancing evaluation cost and domain coverage. Three creativity-psychology experts write each entry and conduct iterative cross-checks. An entry is included only after consensus validation and revision, and all inputs are further screened to remove sensitive information. Details are provided in Appendix A.

Baselines. To compare AlphaContext with both strong LLMs and generation frameworks, we consider baselines from three categories. The prompt template is provided in Appendix I.

(i) **General-purpose LLMs.** We consider

DeepSeek-V3.1, Qwen3-235B-A22B, Llama3.3-70B-Instruct, GPT-5.1, and Gemini-3.0-Pro-Preview as competitive instruction-following models with strong general reasoning and generation capabilities. (ii) **Long-form specialized LLMs.** LongWriter-llama3.1-8b and LongWriter-glm4-9b (Bai et al., 2025) are included as specialized baselines for long-form writing. (iii) **Structured generation frameworks.** DOC-v2 (Yang et al., 2023) combines hierarchical outlining with an adherence controller. CRITICS (Bae and Kim, 2024) performs critic-guided iterative refinement. SS-GEN (Feng et al., 2025) applies constraint-driven hierarchical prompting (STARSOW) for structured story generation. We do not compare with CPIG since it generates short test items and is not designed for long-form context generation.

Evaluation Metrics. We evaluate each generated creativity context along 7 dimensions. Coherence measures narrative consistency, Relevance measures theme alignment, and Engagement measures how motivating the context is for participants. Our evaluation framework is theoretically grounded in Amabile’s Componential Model of Creativity (Amabile, 1983, 2018), which emphasizes task motivation, domain-relevant grounding, and creativity-related processes as core foundations of creative performance. We further evaluate Significance (Okuda et al., 1991; Mumford et al., 2018) to capture real-world relevance and intrinsic task motivation, Concreteness (Guegan et al., 2017) to reflect situational specificity that supports feasible ideation, and Uncertainty (Beghetto and Jaeger, 2022; Beghetto, 2021) to quantify open-endedness that fosters divergent thinking rather than premature closure.

Table 2: Ablation study of AlphaContext on multiple evaluation metrics. Details are the same as Table 1.

Methods	Coherence (↑)%	Relevance (↑)%	Engagement (↑)%	Significance (↑)%	Concreteness (↑)%	Uncertainty (↑)%	Diverse Verbs (↑)%
-w/o HOP	77.96	70.20	76.85	63.55	70.69	76.11	94.25
-w/o MCG	74.38	71.80	72.17	65.76	69.09	71.92	93.79
-w/o ECO	75.62	70.57	71.80	64.53	68.72	70.69	93.36
AlphaContext	81.28	79.06	79.93	71.06	75.49	80.30	96.06

Table 3: Text similarity to expert-designed contexts measured by ROUGE-1, ROUGE-L, and BERTScore. Higher is better.

Methods	ROUGE-1 (↑)%	ROUGE-L (↑)%	BERTScore (↑)%
DeepSeek-V3.1	26.22	20.53	80.94
Qwen3-235B-A22B	20.03	16.42	80.39
Llama3.3-70B-Instruct	20.23	16.32	80.81
LongWriter-llama3.1-8b	19.88	16.33	<u>81.42</u>
LongWriter-glm4-9b	24.42	20.34	81.33
GPT-5.1	15.25	12.46	79.28
Gemini-3.0-Pro-Preview	22.89	18.59	80.31
DOC-v2	18.24	15.65	79.87
CRITICS	17.83	15.14	79.74
SS-GEN	<u>27.80</u>	<u>21.33</u>	81.07
AlphaContext	30.41	25.48	81.88

These psychometric dimensions were iteratively refined through extensive reviews by senior experts in creativity psychology and aligned with established standards from the Future Problem Solving Program (FPSP) (Torrance et al., 1976), which supports strong content validity. Following arena-hard-auto (Li et al., 2025b), we use contexts generated by DeepSeek-V3.1 as the reference baseline and obtain quantified scores through pairwise comparisons with other generated contexts. We also report Diverse Verbs (Fan et al., 2019) to measure action diversity in the context. The prompt template is provided in Appendix I.2. We also report ROUGE-1, ROUGE-L, and BERTScore to measure similarity to expert-designed contexts. Empirically, our real-world human study further supports construct validity by showing significant positive correlations with standardized creativity measures.

Implementation Details. All open-source models are locally deployed and run with vLLM on 8× NVIDIA H200 GPUs. For DOC-v2, CRITICS, and SS-GEN, DeepSeek-V3.1 is used as the generation engine. Additionally, DeepSeek-V3.1 serves as the evaluator model for all judgments. To mitigate position bias in pairwise judgments, we evaluate each context pair twice with swapped orders, and repeat this procedure for two rounds, resulting in four evaluations per pair. We omit standard deviations

in tables since they are consistently small.

5.2 Experimental Results and Analysis

Main Performance Evaluation (To Q1). We compare AlphaContext with 10 baselines on CreaTE to assess multi-dimensional context quality for creativity assessment. Diverse Verbs is computed automatically, while Coherence, Relevance, Engagement, Significance, Concreteness, and Uncertainty are judged by an LLM. We follow arena-hard-auto (Li et al., 2025b) and conduct pairwise comparisons against the DeepSeek-V3.1 output as the reference model. For each subjective metric, we report the positive rate over the reference, which is set at 50% by definition. Table 1 summarizes the results on CreaTE. AlphaContext ranks first on all seven metrics, with the largest gains on Coherence, Engagement, Significance, and Uncertainty, which are key for constructing coherent stimuli that implicitly cue challenges and elicit open-ended creative thinking. Notably, SS-GEN surpasses GPT-5.1 and Gemini-3.0-Pro-Preview on Significance, while AlphaContext further raises it to 71.06%. This suggests that AlphaContext’s planning, search-based generation, and iterative optimization collectively drive consistent improvements across metrics.

Ablation Study (To Q2). To quantify module contributions, we build three ablated variants by removing the HyperTree Outline Planner (HOP), the MCTS-based Context Generator (MCG), and the Evolutionary Context Optimizer (ECO). Table 2 shows that removing any module degrades performance. In particular, removing HOP yields a sharp Relevance drop to 70.20%, indicating that hierarchical planning is critical for keeping assessment cues aligned with the intended theme. Removing MCG lowers Coherence to 74.38% and also reduces Engagement and Uncertainty, suggesting that MCTS search helps preserve long-range structure and the open-endedness needed to elicit creative thinking. Removing ECO decreases all metrics, most notably Uncertainty, showing that MAP-Elites refinement is important for expanding

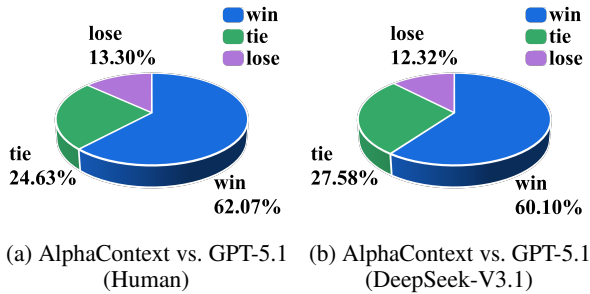


Figure 3: Preference evaluation of AlphaContext vs. GPT-5.1 under human and DeepSeek-V3.1 judgments.

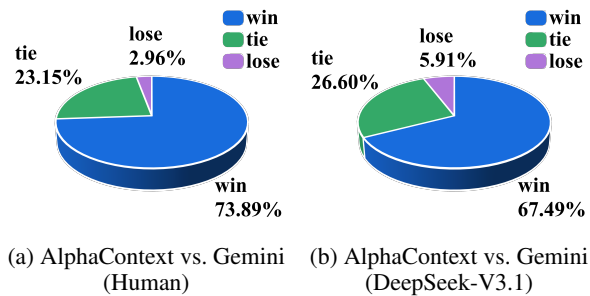


Figure 4: Preference evaluation of AlphaContext vs. Gemini-3.0-Pro-Preview under human and DeepSeek-V3.1 judgments.

stylistic coverage while maintaining assessment cues. This decline arises from two complementary roles of the ECO. First, it expands stylistic coverage by maintaining niche-specific elites in a 3D behavior space, which broadens the diversity of generated contexts. Second, it acts as an effective quality filter through the iterative mutation–evaluation–update loop, which polishes consistency and theme alignment beyond raw MCTS seeds. The ablation results thus validate that the MAP-Elites refinement simultaneously enhances stylistic diversity and core quality while preserving assessment cues. Overall, the ablation results show that the three modules make complementary contributions.

Context Similarity Evaluation (To Q3). To assess how closely AlphaContext outputs match expert-designed contexts given the same inputs, we use 16 expert contexts that were deployed in real creativity assessments and validated by domain experts as references. We compute ROUGE-1, ROUGE-L, and BERTScore, which reflect lexical overlap, long-span matching, and semantic similarity, respectively. Table 3 shows that AlphaContext achieves the best results in all three metrics. It reaches 30.41% ROUGE-1 and 25.48% ROUGE-L, outperforming SS-GEN by 2.61% and 4.15%, and also

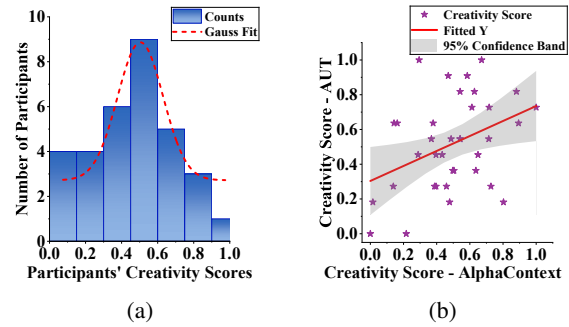


Figure 5: Human study results for AlphaContext. (a): distribution of participant creativity scores with a Gaussian fit. (b): Pearson correlation between AlphaContext-based scores and AUT (Alternative Uses Task) scores.

surpasses strong LLM baselines in BERTScore. Overall, AlphaContext matches the most closely expert-designed contexts, aligning with its goal of generating psychometrically grounded creativity assessment materials. For more details, please refer to our Appendix B.

Preference Evaluation (To Q4). To examine whether the LLM judge aligns with human preferences, we compare AlphaContext with strong baselines (GPT-5.1 and Gemini-3.0-Pro-Preview) via pairwise preference judgments from both human evaluators and DeepSeek-V3.1. To mitigate position bias, each pair is judged twice with swapped orders, and we report win, tie, and lose rates, where win indicates a preference for AlphaContext. Figure 3 and Figure 4 show DeepSeek-V3.1 closely matches human preferences. Against GPT-5.1, AlphaContext wins 62.07% under human judgments and 60.10% under DeepSeek-V3.1, with lose rates around 13%. Against Gemini-3.0-Pro-Preview, AlphaContext wins 73.89% under human judgments and 67.49% with DeepSeek, with low lose rates of 2.96% and 5.91%. The human–LLM agreement is high (Cohen’s $\kappa > 0.8$), supporting the reliability of the LLM judgments for our main evaluations.

Real-World Human Study (To Q5). We validate the assessment effectiveness of AlphaContext-generated contexts through a real-world study with 36 secondary-school students. As shown in Figure 5a, the scores exhibit a symmetric unimodal pattern, and the Gaussian fit closely matches the empirical histogram, suggesting a stable measurement behavior. For criterion validity, we compare AlphaContext-based scores with those of the Alternative Uses Task (AUT), a widely used standardized creativity test. The two scores are collected independently, and the Pearson correlation shows

a significant positive association ($r = 0.3770$), as shown in Figure 5b. Notably, according to standard guidelines on psychology and creativity assessment (Gignac and Szodorai, 2016; Funder and Ozer, 2019; Runco and Acar, 2012; Beaty and Johnson, 2021; Beaty et al., 2022; Benedek et al., 2013), a correlation coefficient of $r = 0.3770$ is regarded as practically meaningful and provides reasonable support for criterion validity. This result indicates that the creativity levels elicited by our generated contexts are consistent with an established benchmark. Overall, the human study provides real-world evidence that AlphaContext can measure student creativity in authentic educational settings. More details can be found in Appendix D.

Case Study (To Q6). To evaluate measurement-level alignment with expert assessment, we conduct a controlled case study on the same theme using three contexts: expert-designed, AlphaContext-generated, and Gemini-3-Pro-Preview-generated. We simulate 30 virtual participants with diverse response styles, collect creativity scores, and compare the induced rankings using Spearman correlation and R^2 fit. Figure 6 shows that AlphaContext better matches the expert context ($r_s = 0.84$) than Gemini ($r_s = 0.58$), indicating closer outcome-level consistency with expert-designed assessments. Details are provided in Appendix E.

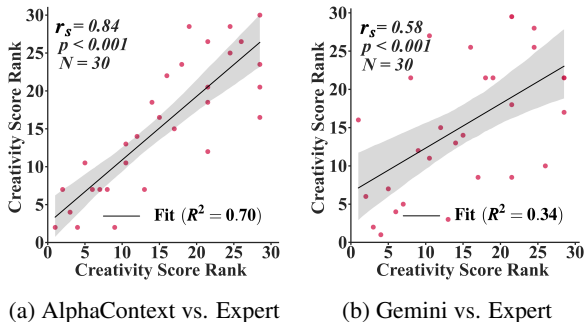


Figure 6: Case study on measurement-level alignment with expert assessment. Scatter plots compare expert-induced score ranks (x-axis) with ranks induced by generated contexts (y-axis) for 30 simulated participants.

Computational Cost Analysis (To Q7). We analyze the computational cost of AlphaContext in terms of generation time and token consumption, in comparison with baseline LLMs and manual expert design. Table 4 reports the average time and token usage required to generate one creativity context. Although AlphaContext requires more tokens and longer inference time than standard zero-shot prompting, this overhead comes from its full

Table 4: Average generation time and token consumption per context.

Method	Time (s)	Tokens (k)
GPT-5.1	23.50	2.36
Gemini-3.0-Pro-Preview	46.31	3.29
AlphaContext (Ours)	226.99	12.89

pipeline and is necessary to ensure psychometric validity beyond what simple prompting can reliably provide. In practice, AlphaContext generates one context in about 6 minutes, making it practical for high-quality dataset construction. By contrast, manual expert design typically requires at least one week (Crabbe, 1989; Barbot et al., 2019a). AlphaContext therefore substantially reduces human labor while maintaining quality. In addition, it uses a locally deployed open-source model, avoiding costly closed-source APIs and improving transparency. Overall, the added computational cost is justified by gains in validity, reliability, and automation.

6 Conclusion

This paper introduces AlphaContext, an evolutionary generator for psychometric assessment contexts that integrates rule-guided outline planning, sentence-level MCTS generation, MAP-Elites quality–diversity optimization, and assessment-guided refinement through virtual participant simulation. Across extensive experiments, AlphaContext consistently outperforms strong LLM baselines and structured frameworks on 7 evaluation dimensions, and shows higher alignment with expert-designed contexts. While it consumes more tokens and requires longer generation time than baseline LLMs, its computational overhead is fully acceptable, and it simultaneously achieves significantly higher assessment validity and generation stability. Human–LLM preference evaluations support reliable automated judging, and a real-world study provides practical validity evidence. Overall, AlphaContext offers a scalable way to produce contexts for creativity assessment while reducing the reliance on scarce expert writing.

AlphaContext is designed as a context generator for human creativity assessment, while also providing standardized contexts for benchmarking LLM creativity. Current experiments focus on future-oriented contexts and a compact expert-curated input set; extending to broader domains, age groups, and languages is an important direction.

Limitations

AlphaContext primarily targets generating psychometrically grounded creativity assessment contexts that are usable in real testing settings. However, achieving stable discourse-level coherence and measurement-relevant cue control relies on sentence-level MCTS and MAP-Elites refinement, which require repeated model calls and scoring. As a result, the overall generation cost depends not only on AlphaContext’s design but also on the underlying LLM and judge configuration, making efficiency comparisons sensitive to the chosen models and evaluation setup. In addition, psychometrically suitable title–theme inputs are still relatively scarce, which can affect the scale of benchmarking and the extent to which conclusions transfer beyond our current future-oriented setting.

AlphaContext offers strong controllability and assessment alignment, while direct prompting is typically cheaper but less reliable for measurement-oriented constraints. Notably, context generation is not a real-time or online task, and the computational overhead is offset by massive reductions in manual expert effort. In future work, we plan to expand expert-curated inputs and use AlphaContext to produce high-quality training data for fine-tuning of lightweight generators. This will improve efficiency while preserving assessment utility, enabling broader coverage across domains, populations, and deployment settings.

Ethical Considerations

This work introduces the CreaTE dataset for creativity-context generation. The title–theme inputs are authored by creativity-psychology experts under a shared specification and contain no personal information. We screen the dataset to remove sensitive or identifiable content, and we prioritize both data quality and ethical compliance during curation. In particular, we apply strict quality-control procedures, including thorough manual review, to ensure broad coverage while proactively addressing potential bias and sensitivity concerns and we curate the dataset in accordance with applicable privacy and research-ethics standards.

Our work also includes a real-world human study to validate the assessment effectiveness of AlphaContext-generated contexts. This study has been reviewed and approved by the Institutional Review Board (IRB) of the affiliated university (IRB Approval No. HR2-0478-2025). Participation was

voluntary, and all participants were informed of the study purpose and procedures, with the right to withdraw at any time without penalty. Before the study, we obtained written informed consent from participants and their guardians, and the consent materials specified the study goals, tasks, potential risks, and data use and protection measures. All collected responses were anonymized by removing personal identifiers, stored securely with restricted access, and reported only in aggregate. The study involved minimal risk, as participants completed open-ended creativity tasks similar to typical classroom activities, and we did not request or record sensitive personal information.

Acknowledgements

We would like to thank the anonymous reviewers for constructive comments. This work is supported by the National Natural Science Foundation of China (No. 62476091), the General Program in Education of the National Social Science Fund of China (No. BEA230071), and the Key Program in Education of the National Social Science Fund of China (No. ABA220028).

References

- T.M. Amabile. 1983. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357.
- T.M. Amabile. 2018. *Creativity in context: Update to the social psychology of creativity*. Routledge.
- Minwook Bae and Hyoungun Kim. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, FL, USA.
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context llms. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore.
- B. Barbot, R.W. Hass, and R. Reiter-Palmon. 2019a. Creativity assessment in psychological research:(re) setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):233.
- Baptiste Barbot, Richard W Hass, and Roni Reiter-Palmon. 2019b. Creativity assessment in psychological research:(re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):233.

- R.E. Beaty and D.R. Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2):757–780.
- R.E. Beaty, D.R. Johnson, D.C. Zeitlein, and B. Forthmann. 2022. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, 34(3):245–260.
- Ronald A Beghetto. 2021. There is no creativity without uncertainty: Dubito ergo creo. *Journal of Creativity*, 31:100005.
- Ronald A Beghetto and Garrett J Jaeger. 2022. Uncertainty: A catalyst for creativity, learning and development. *Creativity Theory and Action in Education*, 6.
- M. Benedek, C. Mühlmann, E. Jauk, and A.C. Neubauer. 2013. Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, 7(4):341.
- A.B. Crabbe. 1989. The future problem solving program. *Educational Leadership*, 7(1):27–29.
- Anne Borland Crabbe. 1982. Creating a brighter future: An update on the future problem solving program. *Journal for the Education of the Gifted*, 5(1):2–11.
- Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy.
- Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, and Dahua Lin. 2025. Creation-mmbench: Assessing context-aware creative intelligence in MLLM. *arXiv preprint*, abs/2503.14478.
- Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen, Guanqun Bi, Minlie Huang, Liping Jing, and Jian Yu. 2025. SS-GEN: A social story generation framework with large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pages 1300–1308, Philadelphia, PA, USA.
- D.C. Funder and D.J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168.
- G.E. Gignac and E.T. Szodorai. 2016. Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102:74–78.
- Jérôme Guegan, Julien Nelson, and Todd Lubart. 2017. The relationship between contextual cues in virtual environments and creative processes. *Cyberpsychology, Behavior, and Social Networking*, 20(3):202–206.
- Jiajun Guo and Sarah Woulfin. 2016. Twenty-first century creativity: An investigation of how the partnership for 21st century instructional framework reflects the principles of creativity. *Roeper Review*, 38(3):153–161.
- Antonio Laverghetta Jr., Simone Luchini, Averie Linnell, Roni Reiter-Palmon, and Roger E. Beaty. 2024. The creative psychometric item generator: a framework for item generation and validation using large language models. In *Proceedings of the 3rd Workshop on Artificial Intelligence and Creativity co-located with 27th European Conference on Artificial Intelligence*, pages 59–73, Santiago de Compostela, Spain.
- Brandon Koh and Angela K-y Leung. 2019. A time for creativity: How future-oriented schemas facilitate creativity. *Journal of Experimental Social Psychology*, 84:103816.
- Hye-Kyung Lee. 2022. Rethinking creativity: Creative industries, ai and everyday creativity. *Media, Culture & Society*, 44(3):601–612.
- Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. Evolving deeper LLM thinking. *arXiv preprint*, arXiv:2501.09891.
- Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, Hamid Alinejad-Rokny, Wei Zhou, and Min Yang. 2025a. STORYTELLER: an enhanced plot-planning framework for coherent and cohesive story generation. In *Findings of the Association for Computational Linguistics*, pages 20818–20846, Vienna, Austria.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, Vancouver, BC, Canada.
- Luchini, Simone A, Maliakkal, Nadine T, DiStefano, Paul V, Laverghetta Jr, Antonio, Patterson, John D, Beaty, Roger E, and Roni Reiter-Palmon. 2025. Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*.
- Aidan Mclaughlin, James Campbell, Anuja Uppuluri, and Yiming Yang. 2024. Aidanbench: Stress-testing language model creativity on open-ended questions. In *NeurIPS 2024 Workshop on Language Gamification*.

- Peter Meusburger. 2009. Milieus of creativity: The role of places, environments, and spatial contexts. *Milieus of creativity: An interdisciplinary approach to spatiality of creativity*, 2:97–153.
- Michael D Mumford, Robert Martin, Samantha Elliott, and Tristan McIntosh. 2018. Creative thinking in the real world. *The nature of human creativity*, pages 147–65.
- Alexander Novikov, Ngân Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint*, arXiv:2506.13131.
- Shawn M Okuda, Mark A Runco, and Dale E Berger. 1991. Creativity and the finding and solving of real-world problems. *Journal of Psychoeducational assessment*, 9(1):45–53.
- Kai Ruan, Xuan Wang, Jixiang Hong, and Hao Sun. 2024. Liveideabench: Evaluating llms’ scientific creativity and idea generation with minimal context. *arXiv e-prints*, pages arXiv–2412.
- M.A. Runco and S. Acar. 2012. Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, 24(1):66–75.
- Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal*, 24(1):92–96.
- Mark A Runco, Burak Turkman, Selcuk Acar, and Mustafa V Nural. 2017. Idea density and the creativity of written works. *Journal of Genius and Emnence*, 2(1):26–31.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can llms generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore.
- Robert J Sternberg. 1984. Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences*, 7(2):269–287.
- Robert J Sternberg and Todd I Lubart. 1999. The concept of creativity: Prospects and paradigms. *Handbook of creativity*, 1(3-15).
- E. P. Torrance, C. B. Bruch, and J. P. Torrance. 1976. Interscholastic futuristic creative problem-solving. *The Journal of Creative Behavior*, 10(2):117–125.
- Tyler J VanderWeele. 2025. Intellectual and viewpoint diversity: Importance, scope and bounds. *Education Sciences*, 15(12):1592.
- Xueyang Wang, Wei Liu, Kaixiang Zhuang, Cheng Liu, Jingyi Zhang, Li Fan, Qunlin Chen, and Jiang Qiu. 2025a. Neural representations of noncentral events during narrative encoding predict subsequent story ending originality. *Science Advances*, 11(17):eadu5251.
- Yixuan Wang, Jiale Feng, Yue Huang, Xuruo Pan, Zhongjing Huang, Zhi Liu, and Hong Qian. 2025b. A style-aware polytomous diagnostic model for individual traits. In *Proceedings of the 28th European Conference on Artificial Intelligence*, pages 2698–2705, Bologna, Italy.
- Hongqiu Wu, Weiqi Wu, Tianyang Xu, Jiameng Zhang, and Hai Zhao. 2025a. Towards enhanced immersion and agency for llm-based interactive drama. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 11166–11182, Vienna, Austria.
- Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025b. Collabllm: From passive responders to active collaborators. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada.
- Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3378–3465, Toronto, Canada.

Appendix

A CreaTE Dataset

AlphaContext takes a title and a theme as input, so evaluation requires inputs that are explicitly designed for creativity assessment. Existing story-generation datasets are ill-suited for this purpose: they target narrative completeness and stylistic richness, but do not ensure that the topic focus, implicit cues, and task intent satisfy psychometric requirements. Using such datasets would confound the evaluation, as poor performance could arise from mismatched inputs rather than the generation method.

To address this gap, we curate CreaTE, a compact yet high-quality input set for creativity-context generation. CreaTE contains 203 title–theme pairs spanning diverse domains, balancing evaluation cost with broad coverage. Each entry is authored by three creativity-psychology experts under a shared specification: the title provides a concrete anchor, while the theme delineates the intended problem space and key creative tensions to be elicited, as shown in Figure 7. A subset of expert-designed

titles and themes is adapted from publicly available FPSP topic resources¹. We refine the dataset through cross-checking and iterative revision, and include an entry only after expert consensus on assessment relevance, clarity, and correctness, with edits applied to remove ambiguity or unintended cues. Finally, we conduct compliance screening to ensure that no sensitive or identifiable information is present.

```
[
  {
    "title": "Youth in Competitive Sports",
    "theme": "Youth competitive sports: performance pressure,
    mental health, injury risk, equity of access, and
    technology-enhanced training."
  },
  {
    "title": "Nanotechnology in Daily Life",
    "theme": "Nanotechnology: smart materials, medicine,
    environmental cleanup, manufacturing change, and long-term
    safety and regulation."
  },
  .....
  {
    "title": "Neurotechnology Future Scene",
    "theme": "Medical Rehabilitation-Focused Neurotechnology:
    Rehab applications, state ethics review, consent, brain data
    security."
  },
]
```

Figure 7: Example input format of CreaTE. Each entry includes a title and a theme, designed to support psychologically grounded creativity context generation.

B Preference Evaluation

We provide additional details on the human evaluation protocol and inter-annotator agreement analysis for the preference study involving Gemini-3.0-Pro-Preview.

For human judgments, we recruited three evaluators and adopted a standardized rating protocol with a pre-study calibration session. The human evaluation checklist is strictly aligned with our metric definitions (Coherence, Relevance, Engagement, Significance, Concreteness, Uncertainty) to ensure consistent interpretation. Inter-rater agreement among the three evaluators meets the required consistency standard, and we report the mean of their judgments as the final human preference result.

Consistent with observations in the main text, DeepSeek’s judgments closely track human preferences, with high human–LLM agreement (Cohen’s $\kappa > 0.8$), further validating the reliability of automated evaluation in our experiments.

¹<https://fpspi.org/topics/>

C Cue Coverage and Diversity in a Case Comparison

Under the same input theme, AlphaContext produces a context fragment with noticeably broader assessment-cue coverage than SS-GEN. Figure 8 shows that AlphaContext can surface multiple challenge dimensions within a single coherent narrative move, so the scenario invites reasoning from several angles rather than focusing on only one. In contrast, SS-GEN typically centers the context around a single dominant cue, which yields a narrower cue footprint and fewer directions for subsequent idea exploration. This qualitative comparison aligns with our design goal: by planning cue placement at the outline level and enforcing outline-grounded generation, AlphaContext increases the diversity of assessment-relevant cues while keeping them implicitly integrated into the story, better supporting creativity assessment that aims to elicit open-ended and multi-perspective thinking.

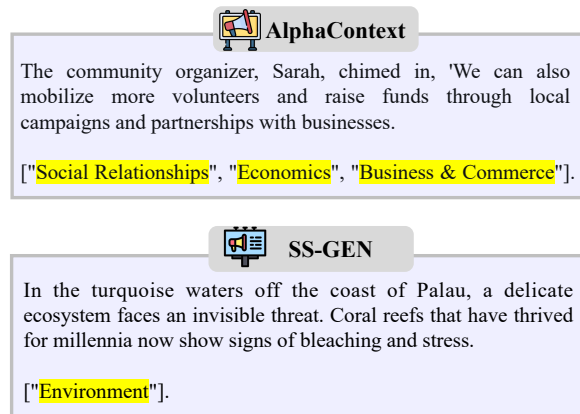


Figure 8: Case comparison under the same input theme: AlphaContext embeds a broader and more diverse set of assessment cues in a coherent fragment, whereas SS-GEN concentrates on a single dominant cue, resulting in narrower cue coverage.

D Detailed Analysis of Human Study

To complement the aggregate Pearson correlation reported in the *Real-World Human Study*, we further examine how specific creativity dimensions align between AlphaContext and a standardized benchmark using Spearman rank correlation. We used Pearson correlation in the main text because the aggregated total scores empirically exhibit an approximately unimodal, near-Gaussian distribution, making Pearson an appropriate summary of linear association. In contrast, dimension-level creativity scores can be more heterogeneous in both

cognitive mechanisms and distributional shapes. We therefore adopt Spearman correlation to provide a rank-based and more robust analysis of whether the two assessments preserve consistent relative ordering at a finer granularity.

Before the formal study, all participants completed the assessment under a standardized administration procedure to ensure comparability across individuals. The instructions, task materials, and time constraints were fixed and delivered in a consistent format. Participants were guided to complete the tasks independently and were discouraged from discussion or external assistance during the session. To protect privacy, we removed personally identifiable information from all collected records, used anonymized participant identifiers for subsequent analysis, and stored the data in an access-restricted manner. Only de-identified responses and scores were used for reporting and correlation analysis.

All human scores used in this analysis were produced under a standardized expert rating protocol. Specifically, three psychology experts served as raters. A domain specialist first developed a detailed scoring rubric and dimension definitions, and then conducted fine-grained training for the raters before annotation. The training session explained each dimension with concrete guidance and calibration examples, ensuring a shared understanding of the scoring criteria and common failure cases. During scoring, the three raters evaluated responses independently. We then computed inter-rater agreement to verify consistency, and the agreement met the required standard. Finally, we report the mean score across the three raters as the human score for each dimension.

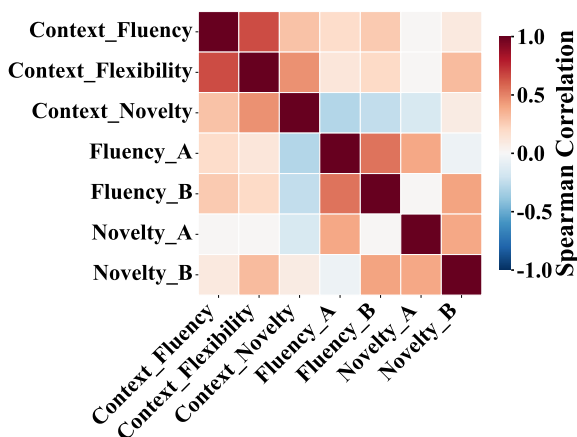


Figure 9: Spearman rank-correlation heatmap between AlphaContext metrics and AUT metrics.

Figure 9 shows the Spearman correlation matrix between AlphaContext dimension scores and Alternative Uses Task (AUT) scores. The AUT metrics are reported for two AUT prompts, denoted by the suffixes `_A` and `_B`. Overall, the heatmap suggests a meaningful but nuanced alignment. In particular, `Context_Fluency` exhibits a consistent positive association with AUT fluency, supporting validity on ideational productivity. Meanwhile, correlations involving `Context_Novelty` are weaker and less directly matched to the standard AUT novelty. This pattern is expected and informative: AUT measures unconstrained divergent thinking, whereas AlphaContext evaluates creativity elicited under explicit semantic and contextual constraints. The results indicate that our method aligns with established benchmarks on general fluency-related signals, while capturing a context-dependent aspect of creativity that may not be fully reflected by standard AUT scoring.

E Extended Measurement-level Alignment Study Across Baselines

To complement the main case study (To Q6), we extend the same measurement-level alignment analysis to all baselines. For each method, we generate a context under the same theme as the expert-designed reference, simulate responses from the same set of 30 virtual participants with diverse response styles, and compute creativity scores under each context. We then compare the outcome-level ranking induced by each method with the expert-induced ranking.

Figure 10 summarizes the alignment results across baselines. We report Spearman rank correlation to quantify ranking consistency with the expert reference, and we also include the corresponding R^2 from a linear fit as a complementary indicator of overall association. Overall, AlphaContext achieves the strongest agreement with the expert reference, suggesting that it best preserves the relative creativity differences elicited by expert-designed measurement at the outcome level. Several long-form or structurally guided baselines show moderate alignment, whereas general-purpose one-shot generators exhibit noticeably weaker agreement. This baseline-wide comparison provides a broader view of measurement consistency and further supports AlphaContext’s advantage beyond single-baseline comparisons.

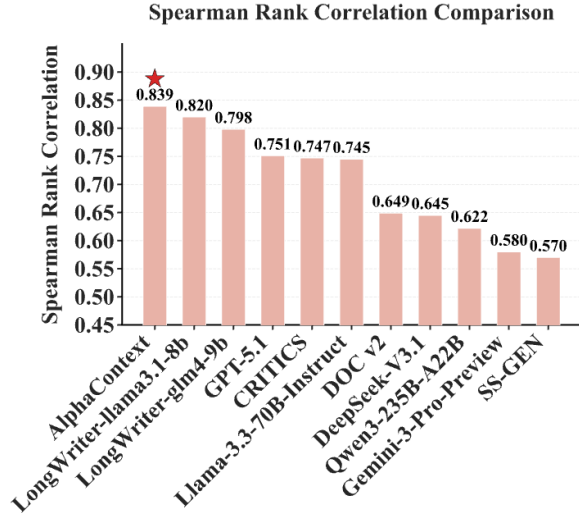


Figure 10: Extended Measurement-level Alignment Study Across Baselines.

F Analysis of MCG Evaluator

This section analyzes the evaluator used in the MCTS-based Context Generator (MCG). MCG formulates long-form context generation as a sentence-level tree search under a planned outline. Each node represents a partial context, and candidate continuations are explored through MCTS. To decide which branches to expand and keep, the MCG relies on an evaluator that scores each newly generated sentence fragment with respect to creativity-assessment context quality.

The evaluator considers three criteria. Cue alignment S_{sc} measures how well the generated sentence matches the outline and embeds assessment-relevant cues. It focuses on faithful adherence to the outline hints, including required themes, keywords, constraints, and intended challenge categories, while encouraging implicit planting of meaningful problem cues such as trade-offs, constraints, stakeholder tensions, and second-order effects, instead of explicitly listing challenges. Imagery vividness S_{im} measures how strongly the text evokes a vivid mental image and immersion through concrete situational and sensory details, including visual impressions, sounds, smells, physical sensations, and emotions, following creativity measurement literature (Wang et al., 2025a). Discourse coherence S_{co} evaluates whether the fragment reads smoothly within the evolving context, with stable entities, natural transitions, and clear causal and temporal continuity.

For evaluation reliability, each criterion is rated on a 5-point Likert scale and then normalized to

$[0, 1]$, where L is the Likert score. The three normalized scores are aggregated with coefficients $\omega_1, \omega_2, \omega_3$, consistent with the main text, to form the scalar evaluation value used by MCG during the search. We examine three coefficient groups to assess sensitivity: Group 1 uses $\omega_1 = \omega_2 = \omega_3 = 0.33$; Group 2 uses $\omega_1 = 0.4, \omega_2 = 0.3, \omega_3 = 0.3$; Group 3 uses $\omega_1 = 0.5, \omega_2 = 0.25, \omega_3 = 0.25$. As shown in Figure 11 and Figure 12, the six subjective metrics exhibit only mild fluctuations across groups, indicating stable evaluation behavior under reasonable coefficient changes. Group 2 yields the highest overall average score, suggesting that moderately emphasizing cue alignment best balances outline-grounded cue placement with vivid writing and coherent discourse. We therefore adopt $\omega_1 = 0.4, \omega_2 = 0.3, \omega_3 = 0.3$ as the default setting for all subsequent experiments.

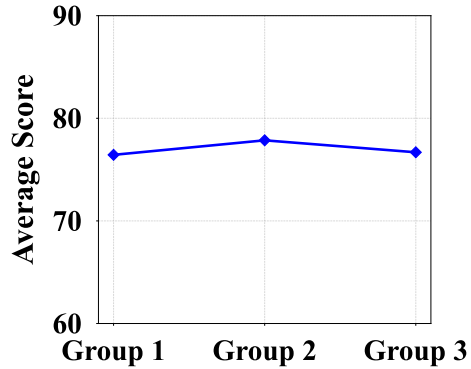


Figure 11: Comparison of the overall average scores across the three weight groups.

G Details of the HyperTree Outline Planner

This section provides additional details for the HyperTree Outline Planner (HTP), complementing the description in Section 4.1. We focus on how HTP operationalizes outline design as hypertree search, where directed hyperedges enable hierarchical divide-and-conquer over discourse structure and assessment-cue placement. Beyond the high-level phases (HT-Select, HT-Expand, HT-Construct, and HT-Decide), we clarify the rule system that governs admissible expansions and explain how LLM-based decisions are integrated to control search scale and semantic validity.

Section G.1 specifies the construction rules of HTP. It includes a static hierarchical skeleton that defines the admissible outline topology, and a set of dynamic, LLM-driven selection rules that prune

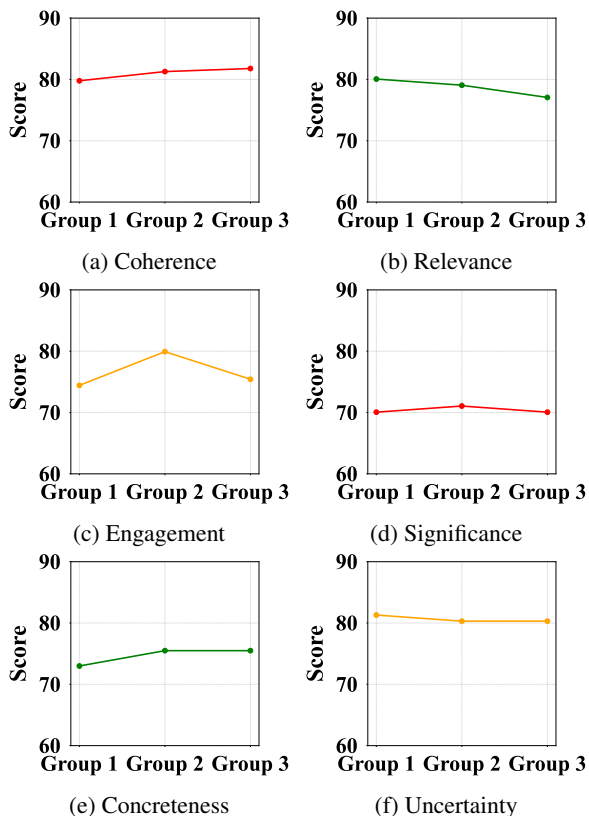


Figure 12: Effect of different MCG evaluator coefficient settings on the six subjective metrics. Subfigures (a)–(f) report the mean scores for Coherence, Relevance, Engagement, Significance, Concreteness, and Uncertainty, respectively.

candidate hyperchains and choose the next divisible leaf node to expand. Section H.4 then presents a fully instantiated outline for the title *AI Partner*, illustrating how these rules concretely materialize into a structured outline suitable for creativity assessment context generation.

G.1 HyperTree Construction Rules

HTP follows a hybrid rule mechanism that combines a predefined context-free grammar with semantic-aware dynamic selection. The context-free grammar provides a static skeleton that constrains the outline to valid hierarchical forms, while dynamic selection rules use an LLM to evaluate candidate hyperchains under a budget and to select the most promising divisible node for expansion. This hybrid design separates structural admissibility from semantic quality control, allowing HTP to maintain global validity while remaining flexible to the input title and theme. The formal definitions of the static skeleton and the dynamic selection nodes are provided in Listing 1.

H Details of the Evolutionary Context Optimizer

This section provides additional implementation details of the Evolutionary Context Optimizer, with a focus on how we operationalize the style-oriented behavior space and how the MAP-Elites archive is constructed and updated. Initialized with the seed contexts produced by the MCTS-based Context Generator, this module searches in natural language space to jointly expand stylistic coverage and improve within-niche quality, yielding a diverse and assessment-ready context archive \mathcal{A} .

H.1 Style-Oriented Behavior Space

To characterize stylistic variations that are highly relevant to future-problem contexts for creativity assessment, we map each candidate context C into a task-specific behavior space B via the descriptor $b(C)$. Following the formulation in Eq. 4, $b(C)$ consists of three interpretable dimensions. The first dimension, $\phi_1(C)$, captures *proximity scope*, measuring how the context is framed from personal daily-life settings to broader public and societal issues (Meusburger, 2009). The second dimension, $\phi_2(C)$, captures *knowledge density*, reflecting how strongly the narrative is grounded in objective evidence such as data cues, mechanistic explanations, and causal constraints (Runco et al., 2017). The third dimension, $\phi_3(C)$, captures *viewpoint diversity*, indicating the breadth of stakeholders represented in the context and whether multi-perspective considerations jointly shape the problem space (VanderWeele, 2025). Together, these three axes form a controllable and interpretable coordinate system to organize stylistic diversity in creativity assessment contexts.

H.2 Descriptor Evaluation, Mutation, and Archive Update

In our implementation, the three behavior dimensions are computed by a fixed-template LLM-based rater, which takes the context text as input and returns normalized scores for $\phi_1(C)$, $\phi_2(C)$, and $\phi_3(C)$ in a structured JSON format. We then discretize the continuous behavior space uniformly to construct a 3D grid archive, where each cell corresponds to a behavioral niche and stores the current elite context.

To enable controllable style shifts in natural language, we implement mutation as a conditional LLM editing process. Each mutation step spec-

```

# ===== Part 1: Fixed Hierarchical Skeleton (Static Rules) ===== 1
2
1. [Plan] -> [Anchor][Scene Setting][Characters & Interaction][Conflict & Challenge][Open Task] 3
   # The plan can be divided into five core narrative aspects.
2. [Anchor] -> [Future Horizon][Place][Scale][Challenge Seeds 1] 4
   # Establishes the fundamental time, space, and scope coordinates.
3. [Scene Setting] -> [Scenario Frame][Constraint Hints][Challenge Seeds 2] 5
   # Defines context and constraints. 6
4. [Characters & Interaction] -> [Interaction Goal][Dispute Focus][Problem Slot][Challenge Seeds 7
   3] 8
   # Constructs interpersonal dynamics. 9
5. [Conflict & Challenge] -> [Challenge Seeds 4][Creativity Triggers] 10
   # Introduces complicating factors. 11
6. [Open Task] -> [Challenge Identification][Solution Exploration] 12
   # Defines the student's objective. 13
14
# ===== Part 2: Dynamic Selection Nodes (LLM-Driven) ===== 15
   # Logic A: When this node is selected for expansion, the LLM selects one option from the 16
   predefined candidate pool based on theme relevance.
9. [Future Horizon] -> {NearFuture (5-15y) | MidFuture | FarFuture | Speculative} 17
10. [Scale] -> {Community | National | International | Space} 18
11. [Scenario Frame] -> {Everyday Life | Urban Infrastructure | Virtual-Reality Fusion | ...} 19
20
12. [Interaction Goal] -> {Co-creation Workshop | Negotiation | Emergency Response | ...} 21
13. [Dispute Focus] -> {Value Conflict | Resource Conflict | Trust Conflict | ...} 22
14. [Creativity Triggers] -> {Uncertainty | Contradiction | Resource Constraints | ...} 23
24
   # Logic B: When this node is selected for expansion, the LLM selects multiple options from 25
   the predefined candidate pool based on theme relevance.
15. [Challenge Seeds 1] -> {{Select 2-3 seeds from Pool}} 26
16. [Challenge Seeds 2] -> {{Select 2-3 seeds from Pool}} 27
17. [Challenge Seeds 3] -> {{Select 3-4 seeds from Pool}} 28
18. [Challenge Seeds 4] -> {{Select 4-5 seeds from Pool}} 29
30
19. [Topic Phrase] -> {{LLM-generated phrase (6-8 words)}} # Summarizes the core conflict based 31
   on Title/Theme. 32
20. [Constraint Hints] -> {{Select 2-3 from: Policy, Budget, Time Limit, Safety, etc.}} # Limits 33
   the solution space.

```

Listing 1: Formal definitions of static rules and dynamic LLM-driven selection rules in the HyperTree Outline Planner.

ifies target values along the three behavior axes and guides the model to revise the context through insertion, deletion, and replacement, so that the candidate moves toward the desired niche while maintaining narrative readability and cue traceability for assessment use. After mutation, each candidate is evaluated in two aspects. Behavior feature evaluation determines its niche assignment via $b(C)$. Quality evaluation is produced by an LLM-based scorer over coherence, relevance, and engagement, and we compute fitness as the uniform average of these three normalized scores to avoid introducing extra hyperparameters. The archive is updated with niche-wise elite replacement: a candidate is inserted if the niche is empty, and otherwise it replaces the current elite only when it achieves higher fitness.

H.3 Iteration Budget

The archive is constructed iteratively. At each iteration, we sample elites from the current archive, generate niche-targeted mutants, evaluate their behavior descriptors and quality scores, and update the archive accordingly. The process terminates when the preset iteration budget is reached, resulting in a context archive that covers diverse stylistic regions while maintaining stable quality within each niche.

H.4 Case Study: Instantiated HyperTree Outline for AI Partner

Listing 2 presents a fully expanded HyperTree outline produced by HTP for the theme *Human-AI companionship and autonomy*. The outline is constructed by iteratively applying the expansion rules in Section G.1, resulting in a valid hypertree whose

branches correspond to alternative hyperchains and whose leaves specify the finest-grained discourse units. Each leaf node is instantiated into a concrete narrative element that can be directly consumed by the downstream context generator, including assessment-relevant cue carriers such as *Trust Conflict* and grounded setting components such as *Everyday Life*. This example illustrates how the rule system yields an explicit, structured outline that supports controlled cue placement while maintaining flexibility in narrative realization.

I Prompt Templates

In this section, we provide the prompt templates used in this study, including the chat template and the evaluation template.

I.1 Chat Template

For DeepSeek-V3.1, Qwen3-235B-A22B, Llama3.3-70B-Instruct, GPT-5.1, Gemini-3.0-Pro-Preview, LongWriter-Llama3.1-8B, and LongWriter-GLM4-9b, we adopt a unified chat template for context generation, as shown in Figure 13. In this template, the system prompt specifies the model’s role and the psychometric constraints required for creativity assessment contexts, whereas the user prompt instantiates the input title and theme and enforces requirements on output formatting and discourse progression.

I.2 Evaluation Template

To reliably evaluate long-form creativity assessment contexts, we adopt a checklist-grounded pairwise judging template, as shown in Figure 14. The judge is instructed to act as an impartial expert in creativity assessment and problem-context design, and is constrained to output a single discrete label. For each subjective dimension in our metric set, *Coherence*, *Relevance*, *Engagement*, *Significance*, *Concreteness*, and *Uncertainty*, we provide a dedicated checklist that operationalizes the criterion as observable properties of the context text. Given two candidate contexts, the judge compares them *only* along the specified dimension and must not introduce any additional criteria.

Concretely, the prompt first specifies the target metric and injects its corresponding checklist, then presents [Context A] and [Context B]. The judge outputs exactly one of five ordered labels, ranging from strongly favoring A to strongly favoring B. This design serves two goals. First, it

mitigates scale drift and instability commonly observed in direct numeric scoring by converting evaluation into calibrated relative comparisons. Second, the per-metric checklist promotes consistency by anchoring judgments to a shared interpretation of each dimension across methods and examples. The resulting labels are subsequently mapped to pairwise comparison outcomes for computing aggregated scores in our evaluation protocol.

J An Illustrative Example of Creativity Assessment Context

This section presents a concrete example of AlphaContext’s generated output for a single input, shown in Figure 15. The input consists of the title *Youth in Competitive Sports* and the theme *Youth competitive sports: performance pressure, mental health, injury risk, equity of access, and technology-enhanced training*. The resulting text is written as an assessment-ready future problem scenario for creativity measurement: it embeds multiple assessment-relevant cues within a coherent narrative, positions the reader as an active problem solver, and keeps the problem space genuinely open-ended by foregrounding plausible trade-offs rather than steering toward a single predetermined solution. Specifically, the context weaves together pressures from competition and external evaluation, potential mental-health and injury risks, unequal access to training resources, and the dual-use role of technology in enhancing performance while introducing new concerns. This example illustrates how AlphaContext operationalizes these design considerations in long-form scenario writing while maintaining narrative clarity and engagement.

Title: AI Partner	1
Theme: Human-AI companionship and autonomy: ethics, emotional reliance, privacy, and governance of pervasive personal AI assistants.	2
Outline Structure:	3
[Plan]	4
[Anchor]	5
[Future Horizon]	6
[NearFuture (5-15 years)]	7
[Place]	8
[City Or Region]	9
[Specific Facility]	10
[Scale]	11
[Community]	12
[Challenge Seeds 1]	13
[Technology]	14
[Ethics & Morality]	15
[Psychological Health]	16
[Scene Setting]	17
[Scenario Frame]	18
[Everyday Life]	19
[Constraint Hints]	20
[Policy]	21
[Budget]	22
[Time Limit]	23
[Challenge Seeds 2]	24
[Technology]	25
[Ethics & Morality]	26
[Characters & Interaction]	27
[Interaction Goal]	28
[Negotiation Meeting]	29
[Dispute Focus]	30
[Trust Conflict]	31
[Problem Slot]	32
AI autonomy in human-AI emotional companionship	33
[Challenge Seeds 3]	34
[Technology]	35
[Ethics & Morality]	36
[Psychological Health]	37
[Conflict & Challenge]	38
[Challenge Seeds 4]	39
[Technology]	40
[Ethics & Morality]	41
[Psychological Health]	42
[Social Relationships]	43
[Law & Justice]	44
[Creativity Triggers]	45
[Uncertainty Cue]	46
[Contradiction Cue]	47
[Open Task]	48
[Challenge Identification]	49
[Prompt student to identify multiple challenges in the scenario]	50
[Solution Exploration]	51
[Prompt student to think of possible response strategies]	52
	53

Listing 2: Instantiated HyperTree outline for the *AI Partner* theme, generated by the HyperTree Outline Planner.

 Chat Template



System Prompt

You are an expert in creativity assessment and problem-context design. Your task is to generate rich, context-based stimulus texts around a specified title and theme in order to elicit students' creative responses. Creative ideation typically relies on prior knowledge and experience; unlimited imagination does not equate to creating something from nothing, and high-quality creativity assessment must be grounded in real, complex, knowledge-supported contexts.

Please provide a multi-dimensional challenge scene that is logically rigorous, thematically focused, and socially meaningful, presented in a concrete, tangible, contextualized narrative.

*The total output length should be approximately ****{1000} words****.*



User Prompt

Please generate a context-based stimulus text based on the title {title} and theme {theme}, strictly following the system-prompt rules.

Requirements:

Naturally embed challenge cues without explicitly naming categories;

Use rich and evocative language that engages all the senses;

Balance the pacing of the narrative;

Do not present conflict or solutions in advance;

Internally follow the logical sequence: Anchor, Scene Setting, Characters and Interaction, Conflict and Challenge, Open Task.

Do not display any section names or labels in the output.

Write as continuous natural paragraphs.

Return only the final, complete generated context.

*The total output length should be approximately ****1000 words****.*

Figure 13: Unified chat prompt template used by baseline LLMs for creativity context generation.



Evaluation Template



BASE_SYSTEM_PROMPT = (

"You are an impartial judge and an expert in creativity assessment and problem-context design. "
"You must follow the user's instructions exactly and only output one of the allowed labels.")



PROMPT_TEMPLATE = (

""""You are asked to evaluate which of the two creativity-assessment problem contexts is better as a stimulus text that can effectively elicit participants' creative responses.

Dimension to evaluate: {metric}

You MUST judge strictly by the following checklist (do not invent extra criteria):

{checklists}

[Context A] {context_a} [Context B] {context_b}

Based on the above, decide which context is better on this dimension.

Output formatting rules (MUST follow):

- Output ONLY ONE label, with NOTHING else.

- Do NOT add any explanation or comment.

- Do NOT repeat or rewrite the contexts.

- The output must be EXACTLY one of:

[[A»B]] # A is significantly better than B

[[A>B]] # A is slightly better than B

[[A=B]] # Tie

[[B>A]] # B is slightly better than A

[[B»A]] # B is significantly better than A

""""

CHECKLIST = {

"Coherence": [

"The problem context is logically consistent and smoothly structured: it clearly anchors time, place, and scope, lets a single central creative challenge emerge naturally from the situation, and links background, events, and constraints without confusing jumps or irrelevant information."

],

"Relevance": [

"The problem context stays tightly aligned with the assigned theme and central creative challenge: most sentences provide specific information, constraints, or stakeholder viewpoints that shape the creative problem space and support people's idea generation, rather than introducing details that have no real impact on how the problem can be understood or solved."

],

"Engagement": [

"The problem context is engaging and motivational: it presents the unresolved creative challenge as meaningful and intriguing, and clearly positions the reader as an active problem solver who is invited to explore possibilities, weigh trade-offs, and imagine alternative solutions."

],

"Significance": [

"The problem context centers on an issue with real-world needs or social significance (e.g., environment, community, technology, health), so that working on it feels valuable and worth caring about."

],

"Concreteness": [

"The problem context clearly locates time, place, people, and relevant circumstances with concrete, meaningful details, so that people can easily form a vivid mental picture of the situation and the constraints and opportunities for creative action."

],

"Uncertainty": [

"The problem context presents a situation and task that are genuinely open-ended and uncertain: multiple viewpoints, formats, and solution paths are possible, and different creative responses could reasonably lead to different future scenarios in real life, rather than implicitly pointing toward a single predetermined textbook-style answer."

],}

Figure 14: Evaluation prompt template for checklist-grounded pairwise judging across subjective metrics.



An Illustrative Example of Creativity Context Generation

Input

Title: *“Youth in Competitive Sports”*

Theme: *Youth competitive sports: performance pressure, mental health, injury risk, equity of access, and technology-enhanced training.*

Output

The predawn air in Millcreek, a community grappling with decades of industrial decline and a 19% youth poverty rate, hung heavy with the scent of rust and rain-soaked concrete as Leo laced his worn trainers. Each stride around the cracked track echoed the fragile promise his family clung to: a college scholarship, their only viable path toward economic mobility in a town where median incomes lag 30% below the state average. Leo hid the persistent ache in his knees---early signs of overuse injury common among adolescent athletes in high-pressure training environments---from his coaches, knowing his performance metrics were now the linchpin not only of his future but of the community's renewed hope.

This hope, however, was increasingly mediated by technology. The Millcreek Youth Athletic Foundation, in partnership with municipal authorities and corporate sponsor Aura Tech Industries, had recently mandated the use of Performa-VR---a mixed-reality training helmet that tracks biometrics and projects ideal movement patterns through its visor. Buried in the parental consent forms, which families signed under time pressure during registration week, was a broadly worded clause granting the city and its corporate affiliates indefinite, transferable rights to all collected physiological and performance data. This clause, as later revealed in a local nonprofit's analysis, enabled the real-time packaging and sale of athlete biometrics to third parties, including sportswear brands and talent agencies.

During training, the helmet's sensors pressed into Leo's temples as he navigated a virtual obstacle course. His actual knees throbbed in protest, even as the system's alerts chastised him for minor deviations from "optimal biomechanical efficiency." A notification flashed: his scholarship eligibility now required maintaining a 95% efficiency rating. Later, as he removed the sweat-drenched headset, Leo noticed a faint tremor in his hands. At the same moment, the system's mental wellness module flagged "elevated anxiety biomarkers," triggering an automated alert to his coaches that threatened his competition status. Unbeknownst to him, that same stress response was already being auctioned as part of a "high-potential athlete" data package.

Matters escalated when a software glitch during a hurdle drill forced Leo's body into an unnatural landing pose, resulting in a sharp pop in his left knee. The system registered it not as an injury but as a "form deviation," uploading the incident to the city's athletic portal where coaches and sponsors could review it. Later that week, while waiting in the gym office, Leo glimpsed his coach's tablet screen: a live bidding interface where corporate partners could purchase exclusive access to athletes' aggregated biometric streams. His own data---including knee stress indicators and anxiety alerts---was tagged as a "premium offering" due to his top regional ranking. Subsequently, a mandatory update rolled out a "mental fortitude" module that monitored real-time pupil dilation and heart-rate variability, deducting eligibility points for signs of stress. For Leo, already concealing his injury and his family's financial precariousness, the result was a rapid decline in his scholarship score with each labored breath.

At home, Leo's mother---a part-time healthcare aide who understood the long-term risks of musculoskeletal injury---placed a hand gently over his knee. Her voice was thick with emotion as she asked, "How can a system that sells your pain as data claim to protect your future, while punishing you for the very injuries it incentivizes you to hide?" Leo's coach, caught between institutional pressure and personal ethics, admitted quietly, "My performance bonus is tied to your compliance metrics. If I report your injury, you lose your scholarship. If I don't, I may be compromising your health. The system traps us all."

The dilemma now extends beyond Leo and his family. School administrators face scrutiny from parent coalitions and education advocates concerned about duty of care. Local public health officials worry about normalized injury patterns in youth athletics. Corporate sponsors, eager to leverage data for talent scouting and product development, operate under minimal regulatory oversight. Journalists from regional outlets have begun filing public records requests about data-sharing agreements between the city and private entities. Even tech ethics advocacy groups are monitoring Millcreek as a potential test case for biometric surveillance in community sports. Leo's hands still tremble as he eyes the helmet. A new notification glows: "Mental fortitude penalty applied---scholarship eligibility now at 72%." The system monetizes his stress in real time, even as it penalizes him for it. His mother's question hangs in the air, unanswered: "Must we really choose between his future and his well-being?"

This scenario illustrates the tensions emerging as technology becomes deeply embedded in youth sports---a space where personal development, economic opportunity, corporate interest, and ethical oversight increasingly intersect. The challenge is not merely technical or individual, but systemic: how can communities like Millcreek harness innovation to support young athletes without commodifying their health or compromising their futures?

Please apply the problem-solving process to analyze technology-enhanced training in youth competitive sports and identify challenges.

Figure 15: Illustrative example of an assessment-ready creativity context generated by AlphaContext, conditioned on a title and a theme prompt.