TRAINING EMERGENT JOINT ASSOCIATIONS: A REINFORCEMENT LEARNING APPROACH TO CREATIVE THINKING IN LANGUAGE MODELS

Mukul Singh
Microsoft, USA
singhmukul@microsoft.com

Ananya Singha
Microsoft, India
ananyasingha@microsoft.com

Aishni Parab
University of California, Los Angeles, USA
aishni@g.ucla.edu

Mansi Uniyal
Microsoft, India
mansiuniyal@microsoft.com

Pronita Mehrotra
MindAntix
pronita@mindantix.com

Sumit Gulwani
Microsoft, USA
sumitg@microsoft.com

ABSTRACT

Associative thinking is the ability to connect seemingly unrelated ideas and is a foundational element of human creativity and problem-solving. This paper explores whether reinforcement learning (RL) guided by associative thinking principles can enhance a model's performance across diverse generative tasks, including story writing, code generation, and chart creation. We introduce a reinforcement learning framework that uses a prompt-based evaluation mechanism, incorporating established divergent thinking metrics from creativity research. A base language model is fine-tuned using this framework to reward outputs demonstrating higher novelty through higher degrees of conceptual connectivity. Interestingly, the experimental results suggest that RL-based associative thinking-trained models not only generate more original and coherent stories but also exhibit improved abstraction and flexibility in tasks such as programming and data visualization. Our findings provide initial evidence that modeling cognitive creativity principles through reinforcement learning can yield more adaptive and generative AI.

1 INTRODUCTION

Large language models (LLMs) are becoming omnipresent in modern AI systems, powering tools for writing, coding, education, design, and decision-making. Their widespread adoption underscores both their impressive generative capacity and their limitations OpenAI et al. (2024) . Despite their scale and fluency, LLMs often struggle to go beyond pattern reproduction, especially when tasks demand creativity, abstraction, or original synthesis. Despite their success in domains like code , math, and reasoning , studies have shown that these models struggle with creative tasks Wilson & Schooler (1991) , producing artifacts that show higher semantic similarity compared to humans Vinchon et al. (2024). The limitation can be explained by the way these models work, where they parameterize the output distribution of tokens conditioned on the sequence of the previous tokens. The model learns to greedily optimize the next token and thus struggles to associate concepts that are spatially distant in the attention maps. Their training objectives prioritize the prediction of the next token over the generation of novel ideas, often favoring coherence over surprise. This is especially more pronounced when the model has not seen those associations before Kaufman et al. (2009).

Associative thinking—a cognitive process that links seemingly unrelated ideas—is widely recognized as a core mechanism underlying creativity Beaty & Kenett (2023); Kaufman et al. (2009). In one of the earliest associative theories of creativity, Sarnoff Mednick Mednick (1962) proposed a hierarchical model of associativity in which people with steep hierarchies generate less creative

responses, while highly creative individuals exhibit flatter hierarchies that enable them to traverse broader semantic distances. The associative process has also been described through an exploration-exploitation frameworkBeaty & Kenett (2023). Individuals switch between exploratory associative searches (finding new categories) and exploitative associative searches (finding elements within a category). Associative thinking has been shown to trigger areas of the brain related to episodic and semantic memory Beaty & Kenett (2023). It enables humans to generate novel insights, metaphors, and solutions by drawing connections across various domains.

Creativity is a desirable trait for any intelligent system. However, current LLMs are not explicitly optimized for creativity. As a result, they do not exhibit the kind of fluid and flexible thinking associated with human creativity. Previous research in psychology and neuroscience has shown that associative thinking is measurable and trainable Wu et al. (2020), offering a compelling blueprint for developing more creative AI.

In this work, we explore whether reinforcement learning (RL) can be used to instill associative thinking in LLMs, thus enhancing their creative capabilities. We design a novel reward function derived from four dimensions of divergent thinking identified in cognitive science literature Beaty & Kenett (2023), and use it to train multiple base language models through RL. The reward signals encourage the model to produce outputs that reflect deeper, more diverse associations, rather than surface-level coherence. We experiment with models of different sizes and varying base levels of creativity Fein et al. (2025).

Our experiments show that models trained with this method outperform baselines in tasks requiring creativity, including story writing, code generation, and chart creation. Qualitative analysis further reveals that RL-enhanced models generate responses that are more imaginative, diverse, and structurally inventive, attributes closely tied to associative thinking.

In summary, we make the following contributions:

1. We introduce a reinforcement learning framework that explicitly promotes associative thinking, enabling language models to generate diverse and conceptually distant ideas.
2. We propose a creativity reward function grounded in cognitive science and show that it is both robust and strongly aligned with human judgments.
3. We demonstrate that fostering associative thinking benefits not only traditionally creative domains such as storytelling but also analytical tasks including coding and data visualization.

2 Related Works

Recent progress in language models has demonstrated the power of explicit reasoning—via techniques like chain-of-thought prompting, scratchpads, and intermediate program synthesis—to solve complex tasks that require multi-step inference OpenAI et al. (2024). These methods generate intermediate "reasoning tokens" that serve both as a decomposition of the problem and as a form of internal scaffolding. While reasoning has been shown to improve accuracy and interpretability across domains like math , code , and logic , most prior work has focused on prompting or output analysis. In contrast, little attention has been paid to how reasoning behavior evolves during training or how it relates to internal representations and skill acquisition.

Additionally, several works have been proposed to constrain the behavior of LLMs to produce more precise systematic outputs Wei et al. (2022); Nye et al. (2021); Yao et al. (2023). These studies focus on convergent thinking, where the goal is to find a single optimal solution. In contrast, creativity involves divergent thinking, where a problem can have multiple solutions.

Creativity, described as the production of useful and original artifacts, has been studied from various perspectives. The 4P model Rhodes (1961) identifies four different dimensions from which creativity can be studied: Person (personality traits and abilities that affect creative capacity), Process (mental and behavioral stages involved in producing creative work), Press (social and cultural environment that fosters or stifles creativity), and Product (creativity of the artifact produced).

Evaluations of creativity from a product perspective include these two primary metrics of originality and usefulness Sternberg & Lubart (1999); Runco & Jaeger (2012); Diedrich et al. (2015) and some-

times include a third factor of style Besemer (1998) or wholeness Henriksen et al. (2015). Assessing creativity typically involves human subject-domain experts, making it less conducive for automated approaches.

An alternative approach, using Guilford metrics Plucker et al. (2010), originally used in the context of Person or Process, has recently been adapted to evaluate AI-generated creative content. The metrics use four core dimensions: fluency (the sheer volume of meaningful ideas produced), flexibility (the diversity of categories within the responses), originality (the uniqueness of answers), and elaboration (the depth or granularity of details within the responses) to estimate overall creativity.

Several studies have directly applied Guilford's creativity framework to assess AI and LLM capabilities across diverse domains. Stevenson et al. (2022) evaluated `GPT-3`'s creativity on the Alternative Uses Test, comparing the model's performance to human psychology students. Recent applications have extended Guilford's metrics beyond traditional domains. DeLorenzo et al. (2024) adapted the four cognitive dimensions to evaluate LLMs within hardware code generation contexts, demonstrating the framework's versatility across technical applications. Similarly, Elgarf et al. (2024) used the four creativity measures to fine-tune `GPT-3` models for collaborative storytelling with children. In this paper, we apply Guilford's metrics to assess overall creativity in the reward function of reinforcement learning.

Unlike previous studies, in this paper, we design a novel reward model to measure associative thinking in a model rollout and use this to train language models, showing that such associative training makes the models better at both conventionally creative (like storytelling) Srivastava et al. (2023)) and non-creative tasks (code generation). Cassano et al. (2023)).

3 METHODOLOGY

We present a framework for training associative thinking in language models through reinforcement learning. Our approach consists of three components: (1) a creativity measurement framework grounded in cognitive science, (2) a composite reward function, and (3) policy optimization.

3.1 PROBLEM FORMULATION

Let $\pi_\theta(y \mid x)$ denote a language model parameterized by θ, generating output y given prompt x. Our goal is to optimize θ such that outputs exhibit enhanced associative thinking—the ability to form meaningful connections between distant concepts—while maintaining coherence.

Given a response y, we extract a set of *associations* $\mathcal{A}(y) = \{a_1, \ldots, a_n\}$, where each association a_i represents a conceptual link between two ideas in the text. The creativity of y is characterized by properties of $\mathcal{A}(y)$.

3.2 MEASURING CREATIVITY

We measure creativity by decomposing each response into semantically meaningful *associations*, treated as atomic "mini ideas." Creativity emerges from the accumulation and interaction of these associations rather than any single element in isolation. We operationalize this notion using four dimensions adapted from Guilford's divergent thinking framework Plucker et al. (2010): *novelty*, *fluency*, *flexibility*, and *elaboration*.

Intuitively, novelty captures how unexpected the associations are, fluency reflects how many distinct associations are generated, flexibility measures the diversity of conceptual domains spanned, and elaboration assesses the depth with which associations are developed. Each dimension is normalized to $[0, 1]$ and contributes to a holistic measure of associative creativity.

We focus on *conceptual* rather than purely lexical associations, allowing abstract or metaphorical mappings to be counted even when not explicitly signposted. Full mathematical definitions, association extraction procedures, and illustrative examples are provided in Appendix .1.

3.3 REWARD FUNCTION

The composite creativity reward aggregates the four dimension scores:

$$R(y \mid x) = \sum_{d \in \{N,F,X,E\}} w_d \cdot s_d(\mathcal{A}(y)) + \beta \cdot R_{\text{coh}}(y, x) \tag{1}$$

where $\mathbf{w} = (w_N, w_F, w_X, w_E)$ are dimension weights satisfying $\sum_d w_d = 1$, and R_{coh} is a coherence term (computed via embedding similarity) ensuring outputs remain contextually appropriate.

3.3.1 LLM-BASED GRADERS

To implement the reward function, we construct automated graders leveraging large language models. Each grader is a checklist-guided evaluator that inspects candidate outputs along the creativity dimensions:

- **Novelty grader**: Prompts the LLM to identify unusual or rare concept combinations within the text.
- **Fluency grader**: Counts distinct associations enumerated by the model.
- **Flexibility grader**: Categorizes associations into semantic clusters and scores their diversity.
- **Elaboration grader**: Requests detailed explanations and scores depth of connections.

We adopt checklist-based graders to reduce variance in free-form evaluations and anchor judgments to interpretable criteria. Unlike holistic scoring, this decomposition allows the reward model to separately capture distinct facets of associative thinking, reducing entanglement between fluency and novelty.

Variance Reduction. To mitigate stochasticity, each grader evaluation is repeated $M = 3$ times with different random seeds, and scores are averaged. Human-annotated samples (50 total) are used to compute per-grader bias offsets, which are subtracted from raw scores to improve alignment with human judgments (achieving Pearson $r > 0.75$).

Preventing Reward Hacking. Graders may occasionally overestimate creativity in verbose but shallow generations. To address this, fluency scores are capped relative to novelty and flexibility, preventing inflation through repetition and ensuring high reward reflects meaningful associative depth.

3.4 REINFORCEMENT LEARNING TRAINING

We fine-tune a base pretrained language model using reinforcement learning to enhance associative thinking. We optimize using Proximal Policy Optimization (PPO) Schulman et al. (2017):

$$\mathcal{L}(\theta) = \mathbb{E}\left[\min\left(\rho_\theta \hat{A}, \text{clip}(\rho_\theta, 1-\epsilon, 1+\epsilon)\hat{A}\right)\right] - \alpha \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \tag{2}$$

where $\rho_\theta = \pi_\theta(y|x)/\pi_{\theta_{\text{old}}}(y|x)$ is the probability ratio, $\hat{A}$ is the advantage estimate, and the KL divergence term regularizes against the base model π_{ref} to prevent catastrophic forgetting and maintain language fluency. We also do separate training using two more RL alorithms; GRPO DeepSeek-AI et al. (2025) and Reinforce Mroueh (2025).

Episodic Training. Training proceeds in episodic fashion, where each episode corresponds to a full model rollout conditioned on a single prompt. Unlike token-level reward shaping, the associative reward is computed only at generation completion, encouraging long-range planning over local token optimization. This design promotes conceptual coherence across distant segments of the output.
Reward Normalization. To stabilize learning, rewards are normalized within each batch of rollouts, and advantage estimates are computed relative to the batch mean. This relative scaling reduces sensitivity to absolute reward magnitude and encourages comparative improvement among samples.
Convergence Monitoring. We monitor convergence via reward stabilization and qualitative improvements in creativity metrics on validation prompts. Early stopping is applied to prevent overfitting and degradation of language fluency.

4 EXPERIMENT SETUP

4.1 TRAINING HARNESS

We conduct experiments to support the learning hypotheses introduced in the paper. We use reinforcement learning as the primary training method since it is closest to task-based knowledge acquisition. We use standard reinforcement learning strategies, including the following: (1) Proximal Policy Optimization (PPO)Schulman et al. (2017); (2) Group Relative Policy Optimization (GRPO) DeepSeek-AI et al. (2025); and (3) REINFORCE Mroueh (2025). We train for 100 iterations with 32 roll-outs per iteration. We train on a cluster of 8xH100, and the overall training takes 2 hours and 35 minutes on average per model.

4.2 BENCHMARKS

We use multiple benchmarks to evaluate both conventionally creative and analytical tasks.

Storytelling: Generating coherent stories from a premise is a creative task even used in high school writing curriculum. We benchmark against the LitBench Fein et al. (2025) benchmark.

Code Generation: Coding is an inherently analytical task involving the composition of mathematical and language constructs. We benchmark against the MultiPL-E Cassano et al. (2023) benchmark, which contains code generation tasks for over 30 high- and low-resource languages.

Visualization: Visualization and data shaping require a mixture of analytical (data pivots, filters, transformations, etc.) and creative (plot choice, colors, elements, etc.) reasoning.

4.3 METRICS

We use different metrics per benchmark to report performance. In particular, For LitBench, we use the proposed LLM-based evaluator, released with the benchmark Fein et al. (2025). For MultiPL-E, we use the test cases that are part of the test-split of the dataset Cassano et al. (2023); whereas for charting, we use the chart quality measurement as proposed by Chakrabarty et al. (2023).

4.4 MODELS AND CONFIGURATION

We report results on both small and large language models—`Deepseek-distill-7B`, `Phi-4-13B`, `GPT-4o-mini`, and `GPT-o4-mini`. We use the default setup of the model with two variations in inference—(1) with reasoning, which is the standard mode, and (2) without reasoning, where the reasoning tokens are turned off and the model behaves like a regular completion model. Appendix C contains the computational cost per model analyzed.

4.5 BASELINE COMPARISONS

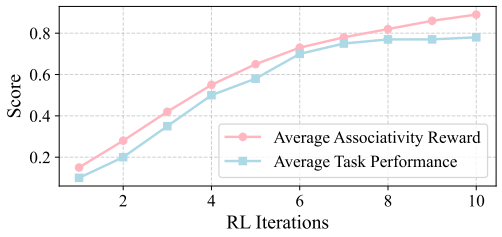
We evaluate our associative thinking RL method against representative baselines spanning pretrained models, prompting strategies, and RL-based adaptations:

Standard Language Models: Base LLM (unmodified pretrained model) and **SFT-Creative**, a model supervisedly fine-tuned on 25K curated creative-writing and problem-solving examples.

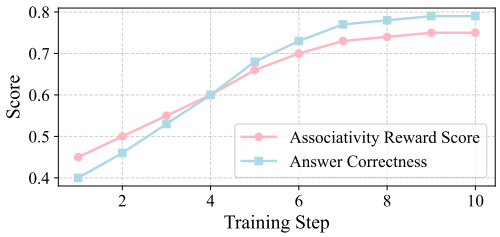
Prompting Strategies: Creative CoT (explicit creative reasoning traces), **Divergent Prompting** (encouraging multiple candidate solutions), and **Role-Play Creative** (persona-based creative prompting).

RL-Based Methods: RLHF-Helpful, trained on helpfulness preferences, and **DPO-Creative**, optimized via creative vs. mundane preference pairs.

Appendix D showcases a +12.1% performance increase of our associative thinking RL against these baseline approaches on Storytelling with `Deepseek-distill-7B` model.



(a) Creativity RL convergence showing alignment between peak associativity reward and task performance.



(b) Stability of creativity reward scores and their alignment with task accuracy over training.

Figure 1: Analysis of training for associative thinking RL models. (a) Convergence of creativity rewards and task performance. (b) Stability of creativity reward and relationship to task accuracy.

Table 1: Performance Improvements from Associative Thinking RL Training (%)

Model	Storytelling	Code Gen.	Viz.
Deepseek-distill-7B	12.1	8.3	9.5
Phi-4-13B	13.4	7.2	10.2
Phi-3.5-instruct	9.8	**-1.5**	8.7

5 Results

We answer the following research questions:

RQ1: Does reinforcing associative thinking improve creativity in language models?

RQ2: Does associative thinking reward captured creativity in generations?

RQ3: Do gains in associative thinking transfer across non-creative and analytical domains?

Appendix B talks about the robustness of the result by analyzing the hyperparameter sensitivity with statistical significance using paired t-tests with Bonferroni correction. Cohen's d, effect size, indicates a mid-to-large effect with a range from 0.62 to 1.24. Additionally, Appendix F provides a qualitative analysis with the generated examples from our associative thinking RL approach.

5.1 RQ1: Creativity Improvements with RL

5.1.1 Improvements in Creativity

We evaluate the improvement in performance on the benchmarks through associative reward training. For this experiment, we evaluate the models on the benchmark in the base setting and compare it with the trained version. Table 1 shows the results across different models.

We can see that the training improves performance across all models and domains (8-13%) indicating that associativity helps the model in problem-solving. Further, the improvement over the storytelling benchmark is the highest, which is not surprising since it's a conventionally creative task. Code generation, being a conventionally analytical task, shows the least improvement but still shows improvement for all models except `Phi-3.5-instruct`, for which it regresses.

We also evaluate the convergence of the training process across RL iterations. Figure 1a shows the performance of the models against increasing RL iterations. We see that the model rewards and task performance peak around the same region, indicating that the model performance gains are aligned to the average associativity reward for its generations.

Figure 2 provides a radar chart visualization comparing the four creativity dimensions across models before and after training, demonstrating consistent improvements across all dimensions with particularly strong gains in novelty and flexibility.

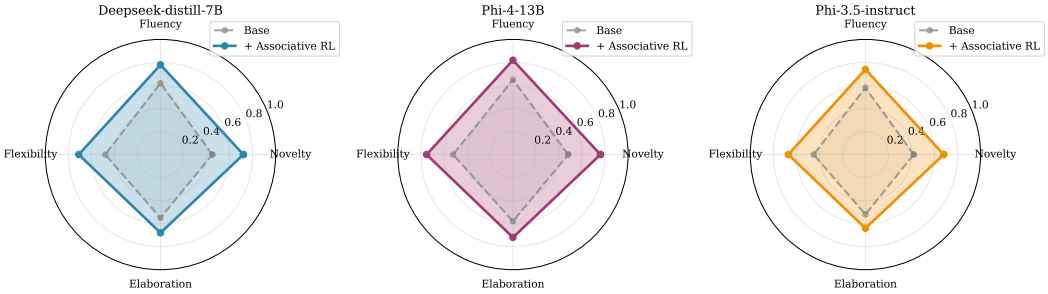
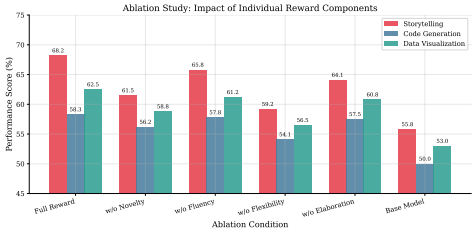


Figure 2: Radar charts showing improvement in creativity dimensions (Novelty, Fluency, Flexibility, Elaboration) for each model. Gray dashed lines indicate base models; colored lines indicate models trained with associative RL.

Table 2: Impact of Removing Individual Reward Components on Performance (%)

Configuration	Storytelling	Code Gen.	Viz.
Full Reward	**68.2**	**58.3**	**62.5**
w/o Novelty	61.5	56.2	58.8
w/o Fluency	65.8	57.8	61.2
w/o Flexibility	59.2	54.1	56.5
w/o Elaboration	64.1	57.5	60.8
Base Model	55.8	50.0	53.0

Figure 3: An ablation study showing the performance impact of removing individual creativity dimensions from the reward function.



5.1.2 Reward Component Importance

To understand the contribution of each component in our associative thinking reward, we conduct systematic ablation experiments. We train variants of our model by removing one creativity dimension at a time from the reward function. Figure 3 visualizes these ablation results, demonstrating the differential contribution of each reward across task types.

Table 2 reveals several important findings. First, all creativity dimensions contribute positively—removing any single component degrades performance. Second, **Flexibility** emerges as the most critical dimension, with its removal causing the largest performance drops across all tasks (9.0% for storytelling, 6.0% for visualization). This aligns with cognitive science research suggesting that the ability to traverse diverse conceptual spaces is fundamental to creative problem-solving. Third, **Novelty** is particularly important for storytelling tasks, while **Elaboration** has a relatively smaller impact, suggesting that depth of explanation is less critical than the breadth and uniqueness of associations.

5.1.3 Creativity Dimensions Across Tasks

We further examine how these creativity dimensions relate to task-specific performance gains. To understand which creativity dimensions most influence specific task types, we compute correlation coefficients between individual dimension scores and downstream task improvements. Figure 4 presents the correlation matrix. Key observations include:

Storytelling: Novelty ($r = 0.82$) and flexibility ($r = 0.78$) show the strongest correlations, suggesting creative narratives benefit most from unexpected ideas and cross-domain connections.
Code Generation: Fluency ($r = 0.62$) shows moderate correlation, while flexibility shows weak association ($r = 0.38$), indicating that analytical tasks prefer focused ideation over conceptual breadth.
Visualization: Elaboration ($r = 0.75$) and Flexibility ($r = 0.72$) dominate, reflecting the importance of detailed design choices and variety in chart creation.

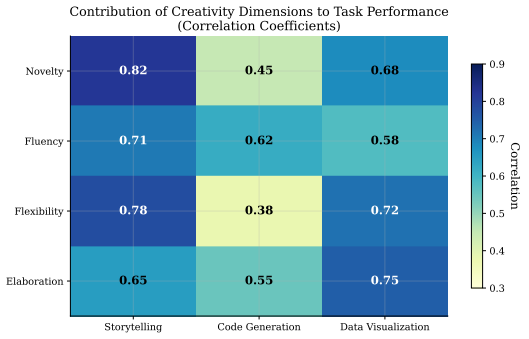


Figure 4: Correlation between creativity dimensions and task performance improvements. Novelty and Flexibility dominate storytelling gains, while elaboration strongly correlates with visual quality.

5.2 RQ2: Creativity Reward Model

5.2.1 Meaningfulness of the Reward Signal

Since we use a non-deterministic LLM-based reward model for evaluating associativity, it is important to assess the robustness and reliability of this reward scoring mechanism. An effective reward function should be (1) consistent across multiple runs, (2) correlated with human judgments of creativity, and (3) predictive of downstream task performance. To evaluate this, we conducted repeated evaluations on a fixed set of generations and measured the variance in reward scores. Figure 1b shows the distribution of reasoning scores and answer correctness as training progresses. The reward scores exhibit low variance, indicating stability despite the underlying stochasticity of the LLM evaluator. Furthermore, we performed correlation analysis between the associative reward and human-rated creativity scores on a sample of story generations. The results showed a strong positive correlation (Pearson's $r = 0.78$), confirming that the reward captures meaningful creative attributes.

Finally, we observe that improvements in associative reward closely track task accuracy, suggesting that the reward effectively guides the model towards more creative yet relevant generations.

5.2.2 Stability and Consistency

Beyond correlation with human judgment, a reliable reward should also exhibit stable and well-behaved score distributions. We analyze how training affects the distribution of associative reward scores across generated outputs.

Figure 8 shows that training produces two key effects: (1) a rightward shift in mean reward scores ($\Delta\mu \approx 0.25 - 0.30$), indicating more creative outputs on average, and (2) a reduction in variance, suggesting that trained models produce consistently creative outputs rather than occasionally high-scoring samples. This consistency is particularly important for practical applications where reliable creative assistance is required.

5.2.3 Learnability across RL methods

We next examine whether the observed reward behavior is robust across different reinforcement learning algorithms. We compare three RL algorithms for training associative thinking: PPO, GRPO, and REINFORCE.

Table 3: Comparison of RL Algorithms for Associative Thinking Training

Algorithm	**Storytelling**	**Code Gen.**	**Viz.**	**Time(h)**
PPO	**67.9**	**58.3**	**62.5**	2.5
GRPO	66.2	57.1	61.8	3.2
REINFORCE	62.8	54.5	58.2	1.8

Table 3 shows that PPO achieves the best performance across all tasks while maintaining reasonable training time. GRPO shows competitive results but requires 28% more training time due to its group-relative optimization. REINFORCE, while the fastest, underperforms due to high variance in gradient estimates—particularly problematic for creativity rewards, which exhibit inherent stochasticity.

5.3 RQ3: Transfer to Non-Creative Domains

5.3.1 Shift from Creative to Analytical tasks

To understand if the benefits of associative thinking extend beyond conventionally creative tasks, we evaluate model performance on analytical tasks such as code generation and data visualization. Table 4 presents results comparing base models with associative reward-trained models on

Table 4: Transfer of Associative Thinking Gains to Analytical Domains (%)

Model	Code Generation	Visualization
Deepseek-distill-7B	7.9	9.5
Phi-4-13B	6.3	10.0
Phi-3.5-instruct	-1.2	8.5

these tasks. We observe consistent improvements on data visualization tasks (up to 10%) indicating that associative thinking aids in creative decision-making like chart design. However, for code generation, results are more mixed. While most models show modest gains, some models (e.g., `Phi-3.5-instruct`) experience slight performance degradation, suggesting that enforcing associative creativity might occasionally conflict with the precision required for programming tasks.

These findings highlight that associative thinking can enhance model flexibility and generalization, but careful balancing is necessary when applying it to highly structured, correctness-critical domains. The observed degradation in certain code-generation settings suggests that unconstrained associative exploration may conflict with the strict syntactic and semantic constraints of programming languages. This highlights the need for task-adaptive reward balancing, where associative incentives are selectively attenuated for correctness-critical domains.

5.3.2 Cross-Task Transfer Learning

To further probe the generality of these gains, we investigate whether associative thinking trained on one task transfers to other domains. For each source task, we train a model exclusively on that task's prompts and evaluate on held-out examples from other tasks.

Table 5: Cross-Task Transfer: Performance Improvement (%) When Training on Source Task and Evaluating on Target Task

Train On	Storytelling	Code Gen.	Viz.
Storytelling	12.1 (direct)	4.2	6.8
Code Gen.	8.5	8.3 (direct)	5.2
Visualization	7.1	3.8	9.5 (direct)
All Tasks	12.1	8.3	9.5

Table 5 reveals that associative thinking skills partially transfer across domains. Training on storytelling provides 35% transfer efficiency to code generation (4.2/12.1) and 56% to visualization (6.8/12.1). Notably, storytelling-trained models transfer better to visualization than code-trained models do, suggesting shared creative reasoning structures between narrative and visual design tasks.

6 CONCLUSION

We demonstrated that training guided by associative thinking principles enhances the creativity of models across multiple domains. Our approach improves performance on storytelling, code generation, and visualization tasks, with strong alignment between the learned reward and human creativity judgments. While benefits extend to analytical tasks, balancing creativity with accuracy remains important. We show the value of combining cognitive insights with RL for AI creativity.

REFERENCES

Roger E. Beaty and Yoed N. Kenett. Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, 27(7):671–683, 2023. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2023.04.004. URL https://www.sciencedirect.com/science/article/pii/S1364661323000943.

Susan P Besemer. Creative product analysis matrix: testing the model structure and a comparison among products–three novel chairs. *Creativity Research Journal*, 11(4):333–346, 1998.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: a scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. Creativeval: Evaluating creativity of llm-based hardware code generation. In *2024 IEEE LLM Aided Design Workshop (LAD)*, pp. 1–5. IEEE, 2024.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C Neubauer. Are creative ideas novel and useful? *Psychology of aesthetics, creativity, and the arts*, 9(1):35, 2015.

Maha Elgarf, Hanan Salam, and Christopher Peters. Fostering children's creativity through llm-driven storytelling with a social robot. *Frontiers in Robotics and AI*, 11:1457429, 2024.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing, 2025. URL https://arxiv.org/abs/2507.00769.

Danah Henriksen, Punya Mishra, and Rohit Mehta. Novel, effective, whole: Toward a new framework for evaluations of creative products. *Journal of Technology and Teacher Education*, 23(3): 455–478, 2015.

Scott Barry Kaufman, Colin G. DeYoung, Jeremy R. Gray, Jamie Brown, and Nicholas Mackintosh. Associative learning predicts intelligence above and beyond working memory and processing speed. *Intelligence*, 37(4):374–382, 2009. ISSN 0160-2896. doi: https://doi.org/10.1016/j.intell.2009.03.004. URL https://www.sciencedirect.com/science/article/pii/S0160289609000300.

Sarnoff Mednick. The associative basis of the creative process. *Psychological review*, 69(3):220, 1962.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification, 2025. URL https://arxiv.org/abs/2503.06639.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi

Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Jonathan A Plucker, Matthew C Makel, and Meihua Qian. Assessment of creativity. *The Cambridge handbook of creativity*, pp. 48–73, 2010.

Mel Rhodes. An analysis of creativity. *The Phi delta kappan*, 42(7):305–310, 1961.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Smita Srivastava, Swati Oberoi, and Vishal K Gupta. The story and the storyteller: Strategic storytelling that gets human attention for entrepreneurs. *Business Horizons*, 66(3):347–358, 2023.

Robert J Sternberg and Todd I Lubart. The concept of creativity: Prospects and paradigms. *Handbook of creativity*, 1(3-15), 1999.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*, 2022.

Florent Vinchon, Valentin Gironnay, and Todd Lubart. Genai creativity in narrative tasks: Exploring new forms of creativity. *Journal of Intelligence*, 12(12):125, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. 35:24824–24837, 2022.

Timothy Wilson and Jonathan Schooler. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology*, 60:181–92, 03 1991. doi: 10.1037//0022-3514.60.2.181.

Ching-Lin Wu, Shih-Yuan Huang, Pei-Zhen Chen, and Hsueh-Chih Chen. A systematic review of creativity-related studies applying the remote associates test from 2000 to 2019. *Frontiers in psychology*, 11:573432, 2020.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

A Extended Methodology

.1 Creativity Scoring and Association Extraction

This appendix provides formal definitions and implementation details for the creativity metrics used in Section 3.2.

.1.1 Associations as Units of Creativity

We treat each semantically coherent conceptual link in a response as an *association*. Associations are not restricted to surface lexical co-occurrence; instead, they capture meaningful conceptual mappings. For example, framing *entropy* as *emotional decay* or describing *recursion* as a personality trait constitutes an association even in the absence of explicit metaphor markers.

Associations are extracted at the clause or phrase level. This granularity balances over-segmentation (which can artificially inflate fluency) against under-segmentation (which may collapse multiple ideas into a single unit).

.1.2 Creativity Dimensions

Each response is scored along four dimensions, all normalized to the range $[0, 1]$.

Novelty (s_N). Novelty measures how unusual an association is relative to typical usage patterns. We operationalize this using normalized pointwise mutual information (PMI) computed over a reference corpus: associations with lower empirical co-occurrence receive higher novelty scores.

Fluency (s_F). Fluency captures the number of distinct associations generated in a response. To discourage verbosity-based inflation, we normalize by response length, ensuring that higher scores reflect genuine idea generation rather than longer outputs.

Flexibility (s_X). Flexibility measures the diversity of semantic categories or conceptual domains spanned by the associations. Responses that traverse multiple domains (e.g., emotional, physical, computational) score higher than those confined to a single thematic space.

Elaboration (s_E). Elaboration assesses the extent to which associations are developed and explained. Rich, meaningful expansion increases the score, while diminishing returns are applied to excessively detailed or repetitive elaboration.

.1.3 Illustrative Example

Consider a story that maps quantum superposition to indecision in human relationships. This mapping contributes to novelty. Further discussion of probabilistic branching or observational collapse increases elaboration. If analogous mappings extend across emotional, physical, and computational domains, flexibility is also enhanced. The total creativity score reflects the combined contribution of all such associations.

.2 Training Algorithm

Algorithm 1 summarizes the complete training procedure.

A Evaluation Dataset Statistics

The table 6 gives an overall summary of the used benchmarks along with its evaluation metrics and source.

Algorithm 1 Associative Thinking RL Training

Input: Base model π_{θ_0}, prompts $\mathcal{D}$, graders $\{G_d\}$, weights $\mathbf{w}$
Output: Trained policy π_θ

Initialize $\pi_\theta \leftarrow \pi_{\theta_0}$, $\pi_{\text{ref}} \leftarrow \pi_{\theta_0}$
for iteration $t = 1, \ldots, T$ **do**
  Sample batch $\{x_i\}_{i=1}^{B} \sim \mathcal{D}$
  for each x_i **do**
    Generate K rollouts: $y_i^{(k)} \sim \pi_\theta(\cdot \mid x_i)$
    Extract associations $\mathcal{A}(y_i^{(k)})$
    Compute rewards $R_i^{(k)}$ via graders and Eq. 1
  end for
  Normalize: $\tilde{R} \leftarrow (R - \mu_R)/\sigma_R$
  Compute advantages $\hat{A}$
  Update θ via PPO (Eq. 2)
  if $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) > \delta_{\text{KL}}$ **then**
    Increase KL penalty α
  end if
end for

Table 6: Summary statistics for evaluation benchmarks.

Task	#	Evaluation Metric	Source
Storytelling	5k	LLM-based creativity score	*LitBench*
Code Gen.	18k+	Test case pass rate	*MultiPL-E*
Viz.	0.5k	LLM-based quality score	ChartEval

B Robustness of the results

B.1 Hyperparameter Sensitivity Analysis

We conduct sensitivity analysis on key hyperparameters to ensure robust results and provide guidance for practitioners.

Learning Rate: We evaluate learning rates in $\{10^{-7}, 5 \times 10^{-7}, 10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$. Performance peaks at 10^{-6}; lower rates, which show underfitting, while higher rates cause training instability and reward hacking.

Rollouts per Iteration: We test $\{8, 16, 32, 64, 128\}$ rollouts. Performance saturates around 32 rollouts with diminishing returns beyond, while training time scales linearly. We select 32 as the optimal trade-off.

Reward Weight Balance: The weights $(\alpha, \beta, \gamma, \delta)$ for (Novelty, Fluency, Flexibility, Elaboration) are tuned via grid search on validation performance. The optimal configuration is $(0.3, 0.2, 0.35, 0.15)$, emphasizing Novelty and Flexibility while maintaining baseline fluency.

B.2 Statistical Significance

We verify the statistical significance of our main results using paired t-tests with Bonferroni correction for multiple comparisons. All reported improvements (base vs. trained) achieve significance $p < 0.001$ except for `Phi-3.5-instruct` on code generation ($p = 0.08$), which shows regression. Effect sizes (Cohen's d) range from 0.62 to 1.24, indicating medium to large effects.

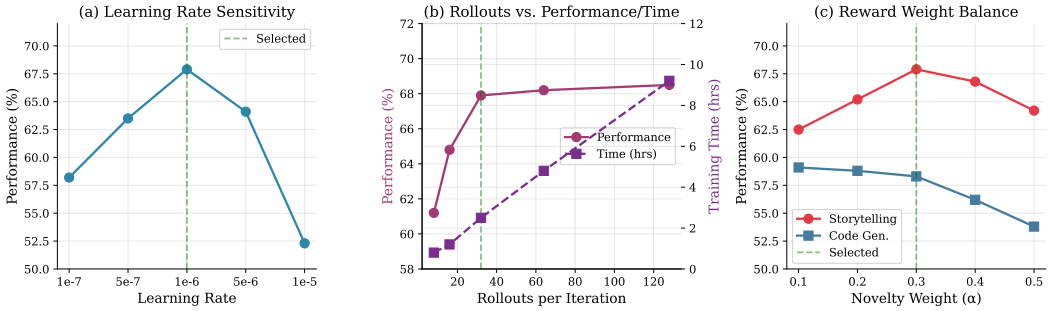


Figure 5: Hyperparameter sensitivity analysis showing (a) learning rate impact, (b) rollouts vs. performance/training time trade-off, and (c) reward weight balance effects on different tasks.

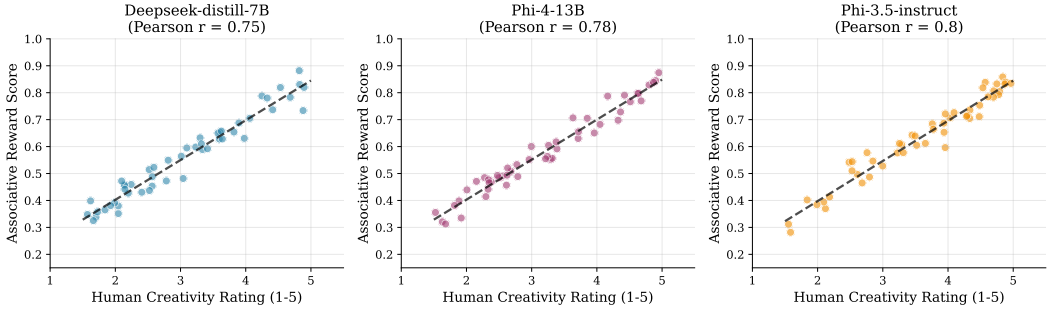


Figure 6: Correlation against human rewards

B.3 Human vs. Associative Reward Alignment

Table 8 reports the alignment between human creativity judgments and associative reward scores across models. All models exhibit strong positive Pearson correlations, indicating that higher reward scores reliably track higher human-rated creativity. Models with higher mean reward scores also tend to achieve higher mean human ratings, suggesting that the associative reward captures salient aspects of human creativity evaluation.

C Computational Cost Analysis

Table 9 summarizes the computational requirements. Training costs are modest compared to full pretraining, making our approach practical for research labs and enterprises. The reward model evaluation adds approximately 15% overhead compared to standard RL training due to the multi-dimensional creativity assessment.

D Comparison against Baseline Methods

We compare our associative thinking RL approach against multiple baseline strategies in Table 10.

Our method outperforms all baselines by a significant margin. Notably, DPO-Creative achieves the second-best performance, but our approach provides a 73% larger improvement (+12.1 vs. +7.0). Prompting strategies provide modest gains but cannot fundamentally alter the model's creative reasoning patterns. SFT-Creative shows that exposure to creative text helps, but RL-based optimization of the creativity objective is substantially more effective.

The figure 8 shows how training affects the distribution of the associative reward scores across generated outputs.

Table 7: Statistical Significance of Performance Improvements

Model	**Task**	p**-value**	**Cohen's** d
Deepseek-7B	Storytelling	< 0.001	1.18
Deepseek-7B	Code Gen.	< 0.001	0.85
Phi-4-13B	Storytelling	< 0.001	1.24
Phi-4-13B	Code Gen.	< 0.001	0.72
Phi-3.5	Storytelling	< 0.001	0.94
Phi-3.5	Code Gen.	0.08	-0.15

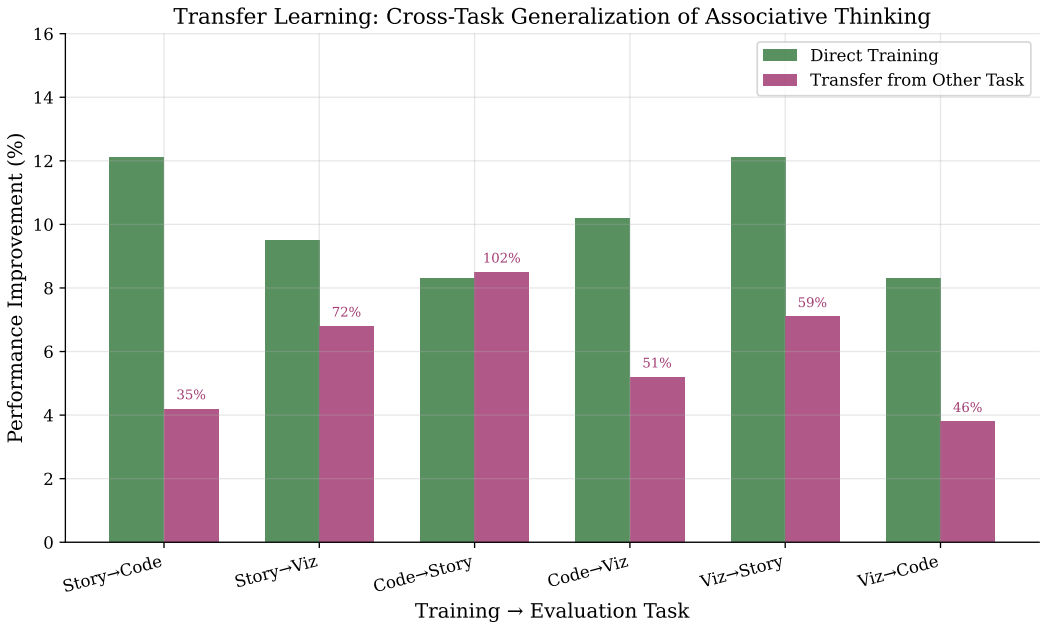


Figure 7: Transfer learning analysis showing how associative thinking gains generalize across tasks. Direct training gains (green) compared to transfer from other tasks (purple).

E Stability and Consistency Extension

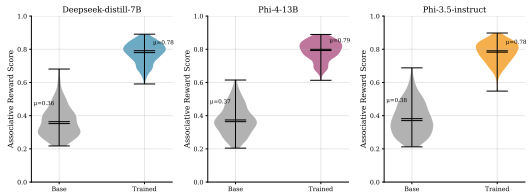


Figure 8: Violin plots showing the distribution of associative reward scores for base models (gray) vs. trained models (colored). Training shifts the distribution toward higher scores with reduced variance.

F Qualitative Analysis

To provide intuition beyond quantitative metrics, we present representative examples comparing base model outputs with our associatively trained model.

Storytelling Prompt: *"Write a short story about a librarian who discovers something unexpected."*

Table 8: Alignment Between Human Creativity Ratings and Associative Reward Scores. Where each represents: Correlation - Pearson's r correlation scores, MRS - Mean Reward Score, and MHR - Mean Human Rating.

Model	Correlation	MRS	MHR
Deepseek-distill-7B	0.75	0.68	3.5
Phi-4-13B	0.78	0.72	3.8
Phi-3.5-instruct	0.80	0.70	3.7

Table 9: Computational Requirements for Associative Thinking Training

Model	GPUs	Time (h)	Memory	Cost ($)
Deepseek-7B	8×H100	2.5	45GB	∼80
Phi-4-13B	8×H100	3.8	68GB	∼120
Phi-3.5-inst	8×H100	2.2	42GB	∼70

Base Model: "Sarah had been a librarian for twenty years. One day, she found an old book behind a shelf. Inside was a map. She followed it and found a hidden room filled with rare books. She was amazed and reported it to the head librarian."

Associative RL (Ours): "Eleanor traced her fingers along the Dewey decimals like a cartographer mapping forgotten territories. Between 133.4 and 133.5—where books on dreams should sleep—she found a volume that breathed. Its pages exhaled memories of readers past: a child's laughter pressed between pages 42 and 43, a widow's tear crystallized in the marginalia. The library, she realized, was not a repository but a living memory palace, and she had become its reluctant archaeologist of souls."

The base model produces a coherent but predictable narrative following common story templates. Our associatively trained model introduces multiple creative elements: the metaphor of Dewey decimals as cartography (cross-domain), the personification of the book breathing (novelty), the concept of crystallized emotions (elaboration), and the philosophical reframe of libraries as memory palaces (flexibility).

Visualization Prompt: *"Create a visualization showing the relationship between coffee consumption and productivity across different professions."*

Base Model: "Create a bar chart with professions on the x-axis and two bars per profession showing coffee cups and productivity score."

Associative RL (Ours): "Design a radial timeline where each profession occupies a spoke, with coffee consumption represented as steam intensity rising from cup icons. Productivity pulses outward as concentric rings—the synchrony between caffeine peaks and productive hours creates interference patterns, revealing which professions ride the coffee wave versus those who drown in it."

The base model suggests a standard visualization approach. Our model proposes a novel radial design with multiple creative mappings: steam intensity for consumption, interference patterns for temporal alignment, and the metaphor of "riding vs. drowning" that adds interpretive depth.

G LIMITATIONS

LLM-based creativity graders The design of the creativity reward function relies on prompt-based evaluations and checklist-style graders, which, although automated and scalable, may introduce biases or overlook subtleties in human creativity. These graders are based on language models themselves, raising questions about circularity and alignment drift, especially when both training and evaluation depend on similar LLM architectures.

Limited domains in evaluation The benchmark coverage, though diverse, is limited to 3 domains—storytelling, code generation, and visualization. These were chosen to span a range of cognitive demands, but they do not encompass other important areas where creativity is critical, such as scientific hypothesis generation, musical composition, or social interaction design.

Table 10: Comparison with Baseline Methods on Storytelling (LitBench)

Method	**Score (%)**	**Δ vs Base**
Base Deepseek-7B	55.8	–
Prompting Strategies		
+ Creative CoT	58.2	+2.4
+ Divergent Prompting	59.5	+3.7
+ Role-Play Creative	57.8	+2.0
Fine-tuning Methods		
SFT-Creative	60.3	+4.5
RLHF-Helpful	58.1	+2.3
DPO-Creative	62.8	+7.0
Associative RL (Ours)	**67.9**	**+12.1**

Reinforcement-only training The reinforcement learning process introduces challenges around stability, sample efficiency, and unintended side effects. Despite reward convergence, we observed occasional degradation in fluency or factual grounding, particularly in code-related outputs. This highlights the need for more robust training strategies that can balance creativity with correctness.

Explore hybrid objectives In the future, we plan to experiment with dynamic trade-off associative reward and task-specific correctness signals, particularly for domains with hard constraints such as formal code synthesis or mathematical proof generation.